

ajer

**THE ALBERTA JOURNAL OF
EDUCATIONAL RESEARCH**

VOLUME XXXVI

NUMBER 1

MARCH 1990

**PUBLISHED BY
THE UNIVERSITY OF ALBERTA • EDMONTON**

THE ALBERTA JOURNAL OF EDUCATIONAL RESEARCH

Published in March, June, September, and December
by the Faculty of Education, University of Alberta

A quarterly journal devoted to the dissemination, criticism, interpretation,
and encouragement of all forms of systematic enquiry into education and
fields related to or associated with education.

EDITOR: Thomas O. Maguire

CONSULTING EDITORS

Rod Clifton
University of Manitoba

Dianne Common
Lakehead University

Carol Crealock
*University of
Western Ontario*

Janet Donald
McGill University

John Hattie
University of Western Australia

Billie Housego
*University of
British Columbia*

Lilian Katz
University of Illinois

Les McLean
*Ontario Institute for Studies
in Education*

John McLeod
University of Saskatchewan

Linda Phillips
*Memorial University of
Newfoundland*

Lauran Sandals
University of Calgary

Keith Sullivan
Dalhousie University

Phil Winne
Simon Fraser University

Ruth Wright
University of Ottawa

FACULTY ADVISORY COMMITTEE

Harvey Zingle (Chair)

Erwin Miklos

Fern Snart

Graham Fishburne

AJER gratefully acknowledges support from the Social Sciences and
Humanities Research Council of Canada and from the Alberta Advisory
Committee for Educational Studies.

The subscription rate is \$32.00 per year for individuals, \$40.00 per year for
institutions. Add \$8.00 for delivery outside Canada. Single copies are \$8.00
each. Please make cheques payable to *The Alberta Journal of Educational
Research*. Back issues are available; rates supplied on request. Claims for
undelivered copies must be received within three months of publication.

Address all communications and manuscript submissions to *The Alberta
Journal of Educational Research*, Faculty of Education, Publication
Services, 4-116 Education North, University of Alberta, Edmonton,
Alberta, Canada, T6G 2G5.

SECOND CLASS MAIL REGISTRATION NUMBER 1436



The Alberta Journal of Educational Research

Volume XXXVI, Number 1

March 1990

ASSESSMENT IN THE CLASSROOM

John O. Anderson (Guest Editor)

EDITORIAL: <i>Assessing Classroom Achievement</i>	1
<i>J.O. Anderson</i>	
Classroom Processes in Evaluating Student Achievement	4
<i>R.J. Wilson</i>	
Administrative Response to Classroom Testing Data: A Problem Solving Perspective	18
<i>P. Nagy and R. Moorhead</i>	
From a Horizontal to a Vertical Method of Integrating Educational Diagnosis with Classroom Assessment	35
<i>C. Dassa</i>	
Measurement and Evaluation Practices of British Columbia Science Teachers	45
<i>D.J. Bateson</i>	
Current Educational Climate in Relation to Testing	52
<i>W.T. Rogers</i>	
Toward Improving Assessment of Students with Special Needs: Expanding the Data Base to Include Classroom Performance	65
<i>D.G. Bachor</i>	
Time to Replace the Classroom Test with Authentic Measurement	78
<i>L.D. McLean</i>	
Assessment in the Classroom: What is the Role of Research?	85
<i>R.E. Traub</i>	
Toward a Relevant Classroom Assessment Research Agenda	92
<i>R.J. Stiggins</i>	

ACKNOWLEDGMENT

The review process is central to the quality of this journal. We thank the following people who reviewed manuscripts in 1989. Their thoughtful critiques contribute to the advancement of theory, research, and practice in education.

Allison, D.
Anderson, B.
Anderson, J.
Assheton-Smith, M.
Bacchus, M.K.
Bachor, D.G.
Boak, T.
Boersma, F.
Bowd, A.
Brown, A.
Burnett, D.
Callan, E.
Campbell, R.
Connelly, M.
DeBlois, C.
Donald, J.
Drost, D.
Dupuy-Walker, L.
Fagan, W.T.
Friesen, D.
Fris, J.
Froese, V.
Greene, M.
Hartman, B.
Haughey, M.
Holmes, B.
Housego, B.

Ivany, G.
Jacknicke, K.G.
Kass, H.
Kysela, G.M.
Laine, C.
Lam, J.
Lang, L.
Lucas, B.
Lupart, J.
MacKay, A.
Marshall, D.
Massey, D.
McIntosh, R.G.
McLean, L.
Miklos, E.
Miller, R.
Misanchuk, E.
Mokosh, E.
Oberg, A.
O'Neill, P.
O'Reilly, R.
Paquette, J.
Patterson, R.S.
Platter, E.
Pugh, R.
Racette, G.
Ragsdale, R.

Ratsoy, E.
Riffel, T.
Robitaille, D.
Rogers, T.
Ryan, B.
Sackney, L.
Samiroden, W.D.
Sande, D.J.
Sawada, D.
Shore, B.
Sitko, M.
Skau, K.
Snart, F.
Sobsey, R.
Tardif, C.
Taylor, G.
Thomas, A.
Timmons, B.
Uhl, N.
Urion, C.
Wasserman, S.
Wilde, W.
Wilgosh, L.
Worth, W.H.
Yore, L.

EDITORIAL

Assessing Classroom Achievement

John O. Anderson

University of Victoria

There is increased interest in the relationship between educational measurement and the classroom practices used in evaluating student achievement (Stiggins, 1988). Questions have been raised as to the utility of procedures developed through educational measurement for teaching and learning.

Although the testing of student achievement is common practice in educational settings, only recently have researchers begun to study classroom testing in distinction from psychometrics. Educational measurement specialists have attempted to bridge the gap between psychometrics and the measurement of educational achievement in the classroom. The educational measurement literature addresses issues such as test reliability, validity, standard error of measurement, item bias, test equating, models underlying response data, procedures for test design and construction, item analysis, test scoring and analysis, and standard setting. But this work is based primarily on large-scale test programs in which the test is administered, analyzed, and interpreted as a single, isolated entity such as a final examination, a system-wide assessment, or a commercially developed standardized test. In contrast, the assessment of student achievement by teachers in classrooms is characterized by multiple testings or data collections throughout an instructional period using an eclectic array of information generating, collecting, and collating procedures (Anderson, 1989).

This special issue of *The Alberta Journal of Educational Research* contains papers presented to the Conference on Classroom Testing held at the University of Victoria in February 1989. The focus of the conference was classroom testing: What testing is going on in classrooms? What is the climate of educational testing? What could be going on, and what should be going on in classrooms in regard to testing the educational achievement of students? The intent of the conference was to bring researchers together to consolidate some of the major elements of this emergent field and articulate directions for future research initiatives. The conference was jointly supported by funding from the Social Sciences and Humanities Research Council of Canada and the British Columbia Ministry of Education.

The participants are active researchers in the field of classroom testing who are making significant contributions to the development of classroom testing as a distinct measurement activity and research field. The papers are organized into the following sections.

Current Research on Classroom Testing

Robert J. Wilson of Queen's University reported the findings to date of the *Ecology of Evaluation: Evaluation within Educational Settings*, a project examining the student evaluation practices of classroom teachers. The study involves a number of schools throughout Canada in an intensive descriptive study of the methods and instruments used in the evaluation of student achievement. Philip Nagy of the Ontario Institute for Studies in Education and Roslyn Moorhead of The University of Western Ontario reported on the results of a systematic analysis of administrator response to the results of an investigation into classroom testing practices. The study analyzed the problem solving behavior of school principals in regard to testing situations in a school setting. Further, administrator responses to discrepancies in classroom testing practices were also reported. Clement Dassa of the Université de Montréal reported on the project *Elaboration d'un modèle d'intégration du diagnostic pédagogique aux apprentissages scolaires*. The project is being conducted on site in a school district in Quebec by a team of researchers at the Université de Montréal. David Bateson of The University of British Columbia reported on the testing and grading practices of British Columbia science teachers. The data were derived from a survey of approximately 1000 grade 4, 7, and 10 science teachers throughout British Columbia. The major issues addressed were the sources of grading information used in evaluating student performance, the connection of grading information to curricular emphasis, the emphasis placed on different types of measurement procedures and instruments, and the patterns of distributions of grades awarded to students.

Current Educational Climate in Relation to Testing

Todd Rogers of the University of Alberta described and discussed the findings of three major investigations into perceptions of public education by students, teachers, parents, trustees, ministry officials, and politicians. Particular attention was paid to those facets with relevance to student evaluation. The investigations involved province-wide surveys of public perceptions of the provincial educational system: *Let's Talk About Schools* (1984, a British Columbia survey of public attitudes toward the provincial educational system) and *The Royal Commission on Education* (1988). Rogers was a key investigator in both of these studies.

What Could Be Going On: Available Technologies and Procedures for Classroom Testing

Dan Bachor of the University of Victoria described a set of procedures which use students' daily classroom performance as the basis of assessment. These procedures allow the regular classroom teacher to determine if a concept has been learned. They also provide the special education teacher with a technically adequate sample of classroom performance that can be translated directly into instructional objectives. Leslie McLean of the Ontario Institute for Studies in Education developed the theme that specialists in educational measurement have given too much attention to the technical aspects of

measurement (e.g., reliability, item analysis, item response models) and too little attention to the educational aspects (e.g., motivation, contribution to teaching and learning). A broader concept of assessment of achievement is advocated, building on work now well advanced in the United Kingdom under the heading "Portfolios" or "Records of achievement." The record of achievement provides a basis for grading while enhancing motivation and documenting attainment comprehensively for parents and employers. Tests can be part of the record, but they play very much a supporting role. Problems of standards and comparability are addressed.

Consolidation and Research Directions

Ross Traub of the Ontario Institute for Studies in Education and Richard Stiggins of the Northwest Regional Educational Laboratory independently provided analyses and syntheses of the papers presented at this conference with a view to identifying researchable problems in classroom testing and student evaluation, and description and discussion of emergent research trends.

References

- Anderson, J.O. (1989). Evaluation of student achievement: Teacher practices and educational measurement. *Alberta Journal of Educational Research*, 35, 123-133.
- Stiggins, R.J. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan*, 65, 363-368.

ROBERT J. WILSON

Queen's University

Classroom Processes in Evaluating Student Achievement

The purpose of this study was to determine whether viewing classroom processes in evaluating student achievement from a decision making model would account for the observed practices. Forty-nine classrooms participated in the study. The classrooms ranged from Kindergarten to Grade 13 and were situated in Ontario and British Columbia. Teachers collected all the instruments they used with one class for one subject over the course of a reporting period. They also completed a checklist for each instrument and a log for the entire process. In addition, an interview was held with each teacher at the end of the reporting period. The results tended to show that teachers beyond the early primary grades use their evaluations mainly to generate marks for reporting purposes. These reports typically reaffirm judgments about achievement rather than develop them, with the result that very little alteration of instruction is made as a result of the evaluation activity.

One approach to describing classroom practices in evaluating students could be to determine the degree to which these practices match those engaged in or approved of by professional evaluators. If a gap is found, then ameliorative action springs fairly readily from the analysis. This is the model underlying Frisbie and Friedman's (1987) suggestions concerning the applicability of the Test Standards as a model for preservice instruction and was also the assumption underlying Mayo's (1964) paper in the inaugural issue of the *Journal of Educational Measurement*.

What researchers like Gullickson (1984) and Stiggins and Bridgeford (1985) have been discovering in their analyses of classroom practice is that the world of the classroom teacher seems so different from that of the professional evaluator that the gap, far from being narrowed by decades of preservice and inservice instruction, is as wide as ever.

One possible way to reconcile these approaches would be to see both enterprises as exercises in informed decision making. A model based on this construct would see both classroom-based evaluators and professional test-makers as engaging in the activities of helping to choose among possible courses of action through the provision of timely information to inform judgment.

Such a model is worked out in more detail in Wilson and Rees (in press) but will be described briefly here. It assumes that evaluation can be conceived of as three interrelated processes: measurement, judgment, and decision making. Measurement is defined, traditionally, as the assigning of numbers to events according to rules. A watch is a measuring instrument because the rules for defining units, for calibration, and so on are known and accepted

Robert Wilson is a professor of education. He is interested in measurement and evaluation.

widely. Judgment is defined as the conceptual act of matching the information gleaned from the measurement with a set of expectations, a test of fit. Thus we check the watch, not primarily to determine the time, but to compare the information it gives us to various mental benchmarks. If the information is that it is earlier than 10:00, then we refrain from going for coffee, but if it is after 10:00, we go. These benchmarks constitute decision points. Thus the third aspect of the model, decision making, requires a selection of a course or courses of action, derived from the measurement-bound judgments that activate the decision.

The criteria for these elements in the decision making system emerge fairly directly from the overall purpose. The goal of measurement is to provide accurate information under certain prescribed conditions, and the criterion for accuracy of measurement is reliability. Much effort is expended in professional test development to enhance the degree of reliability instruments might possess in various situations. Judgments, however, take the issue partially out of the technical realm by introducing the concept of use. The degree to which an instrument's results can help in forming adequate judgments about a situation is a measure of the instrument's validity. As various authorities remind us, validity is in the hands of the user; although the test maker may anticipate and recommend certain appropriate uses of a test, the client is the ultimate arbiter of validity. Finally, decisions made as a result of the testing and its interpretation may be seen as helpful, harmful, or neutral. The criterion of utility can be invoked here as relevant for determining whether the courses of action followed as a result of the measurement activity constitute a net gain on those that might have been followed anyway or those that might have been followed with a different measure.

The application of this approach to the work of the professional evaluator can be quite linear and straightforward. Most test makers define a small number of discrete purposes for an instrument, work at creating an acceptable level of objectivity for it, and provide evidence that judgments which flow from an adequate interpretation of certain patterns of results will be better than would be the case without the measure. Because the professional has control over the measurement, most attention will be paid to its characteristics. As the test moves into its judgment and decision making phases, into use, the ability of the professional to ensure validity and utility shrink although responsibility for contributing to them does not. The appropriateness of the decision making approach to classroom evaluation, however, is less direct. Singleness of purpose, so helpful in aiding in the technical sophistication of the professional, is usually absent in the classroom. The teacher as evaluator serves many levels in a system (students, department heads, principals, superintendents, and Ministry officials) as well as some outside the formal system (parents, colleges and universities, employers). All these groups have insistent and frequently contradictory uses for the evaluation information a teacher may generate. Consequently, the tidiness of aim, so necessary for the development of objective measurement, is replaced with a variety of purposes that renders the teacher's decision making process far more complex (Wilson, 1989; Rees, 1989).

Purpose

The research to be reported here (which is part of a larger program conducted by Wilson and Rees) had as its overriding aim the charting of the policy and procedural frameworks for decision making in which teachers and others in the levels described above make their measures and judgments. As part of the analysis of that system, the classroom level and its policies and procedures constituted an important element. In this report, the classroom processes in evaluation are the focus, and only passing references are made to the other aspects of the investigation. The question this paper addresses is: To what extent does a decision making model account for the practice of student evaluation as designed and conducted by teachers in classrooms?

Method

Design

Because of the overall purpose of the research, a multisite, nonrandom sample of classrooms was selected for this study. Two provinces, British Columbia and Ontario, constituted the primary sampling unit. These provinces differ markedly in the way they create, disseminate, and manage educational policy, with British Columbia operating from a more top-down approach and Ontario from a more differentiated policy framework. Two districts in British Columbia and two boards in Ontario constituted the next level with variation in student enrollment between the districts and boards being the primary selection variable. In each of these units, two elementary and two secondary schools were selected. In each elementary school, one primary class and one junior (Ontario) or intermediate (British Columbia) class was selected. In each secondary school, four classes representing a variety of subject matters and departments were selected.

As a result of this process, which included the opportunity for self-selection at the school level, a final sample of 49 classrooms was obtained. The obtained sample contained two anomalies. Two classes in a secondary school in British Columbia produced inadequate data for inclusion, and in the same province, an oversubscription of participants at the elementary level produced an increase of three classrooms in that portion of the sample.

Table 1 indicates the grade level frequencies participating in the study. As British Columbia does not have Grade 13 in its system, this grade is not represented there. In other respects, the obtained sample provides a fairly complete range of classes in the public systems of both provinces, but one that is obviously quite restricted.

Characteristics of the Classrooms

The elementary teachers who agreed to participate in the study were allowed to select one of the subjects they taught for data collection purposes. The secondary teachers were categorized by various departments to ensure a broad coverage of subject specialties. Table 2 indicates the subject matter that was ultimately obtained for the data collection exercise described below. The designation "integrated" accommodated those primary teachers who do not teach subjects discretely. Two of the teachers monitored individual stu-

Table 1
Grade Levels in the Study, by Province

Grade	K	1	2	3	4	5	6	7	8	9	10	11	12
British Columbia	1	1	2	1	1	3	2	0	4	2	0	4	4
Ontario	1	1	1	1	1	0	2	0	0	3	3	4	2
Total	2	2	3	2	2	3	4	0	4	5	3	8	6

Note: Ontario also had 4 grade 13 classes and 1 special class.

dents rather than individual subjects because of the way in which their curriculum and instruction was organized. The only imbalances in subject matter between the two provinces occur in elementary science where three teachers in British Columbia monitored science classes while none did in Ontario, and at the secondary level, the arts in Ontario were more heavily represented in the obtained sample than was the case in British Columbia.

Instrumentation and Data Collection

Once the teachers in the study had been identified by their principals or department heads, they were contacted by one of the two principal investigators and informed of the purposes and procedures of the study.

Each teacher was required to choose a reporting period for a particular subject and to collect all the evaluation instruments used with that class for that period. An instrument was defined as “any test, rating scale, project, checklist, or assignment used to assess student performance.” At the elementary level, a reporting period was usually interpreted to mean the time between the issuance of report cards, whereas at the secondary level, it was usually interpreted to mean a semester or a term. Teachers were given the opportunity to define reporting period in a way that was congruent with their teaching.

Table 2
Subjects Represented by Province and by Level

Level	British Columbia	Ontario	Total
<i>Elementary</i>			
English	2	2	4
Maths	5	4	9
Science	3	0	3
Integrated	1	1	2
Special Education	0	1	1
<i>Secondary</i>			
Humanities	3	3	6
Maths	2	2	4
Science	3	4	7
Social Studies	5	4	9
Arts & Physical Education	1	3	4
Totals	25	24	49

Two data collection instruments were given to the teachers for completion. The first was a log on which the date, instrument description, number of students completing each instrument, number of A's, number of F's, average, and total marks were recorded. Definitions for these categories were stated on the log itself for reference. The number of A's, for example, was defined as "the number of students receiving the highest grade designation" on the instrument. The number of F's was defined as "the number of students who did not receive a passing grade." The average could be given as either a mean or a median.

The second data collection instrument was a checklist that was appended to each instrument. It required the teacher to indicate which of 10 common purposes this instrument was intended to serve, how well the students had performed on the instrument in the teacher's judgment, what measurement characteristics the instrument possessed (written in nontechnical language), and where the instrument had originated. The checklist was completed or attempted for 776 instruments.

At the conclusion of the data gathering period, each teacher was interviewed in a structured format. Interviews averaged about 45 minutes and were conducted one-on-one with either one of the two principal investigators or a research associate. Recordings of the interviews were not made but detailed notes were kept and subsequently transcribed. The purpose of the interview was to explore the policies and procedures that teachers followed in their evaluation activities. Questions dealing with informing their students of the expectations of the course or unit, the preparation and construction of the measures used, types of items favored, changes in evaluation over the time period, use of administered instruments by students and teachers, and sources of guidance for evaluation activities were included in the interview format.

Data Entry and Analyses

The data from the logs and checklists were quantified and analyzed using descriptive statistics, cross-tabulations, and correlations. The data from the interviews were transcribed from notes, returned to the participants for reliability and validity checks, and their amended responses incorporated into the recorded notes. Five of the 49 teachers participating returned amended copies of the interview summaries.

Obviously, no legitimate attempt can be made from these data to make generalizations across classes, departments, schools, districts or boards, or provinces. Nevertheless, they may be suggestive of trends and approaches that could lend themselves to a more systematic sampling plan and a more precise definition of variables in the future. It is in that exploratory vein that the following report of the data and their interpretations is made.

Results

Characteristics of the Measurement

Frequency. Table 3 displays the frequency with which the teachers reported administering instruments to their selected classes. Direct comparisons among teachers are clouded by the understanding that each teacher could

Table 3
Mean Number of Instruments per Class, by Province, Level, and Grade

Grade	British Columbia	Ontario
<i>Elementary</i>		
Kindergarten	34	4
1	12	7
2	21.5	8
3	10	14
4	17	17
5	9.3	
6	15	38
Special		56
Elementary Means	15.8	22.7
<i>Secondary</i>		
8	14.5	
9	11.5	19.3
10		15.3
11	13.5	16
12	10.75	15.5
13		9.5
Secondary Means	12.7	14.8
Total Means	14.2	17.3

decide on the time period for reporting results. Using the concept of the reporting period, however, leaves the impression that students in both provinces would be assessed between 14 and 17 times on average in one course for one reporting period. This is consistent with the findings of another study (Wilson, 1989) that documented the frequency of evaluations for a secondary school in Ontario for a single semester. If any trend is discernible in the data, it maybe that there is a tendency in British Columbia to be fairly consistent across grade levels in terms of the number of evaluations conducted, but that generally in the Ontario sample, there is a tendency to measure somewhat more frequently than is the case in British Columbia.

Another aspect of evaluation frequency is its regularity. The most common pattern of evaluation indicated a fairly equal spacing of the evaluations over the course of the reporting period. Some logs indicated that the evaluations became more heavily weighted over the term, but many did not. Some teachers reported increasing the number of evaluations near the end of the reporting period as report card time approached (at the elementary level in both provinces).

At the secondary level, the suggestion that the subject matter might lend itself to a simple-to-complex rather than a regular pattern of assessment was examined. Such a pattern did not emerge from the interview and log data. Whereas some social studies teachers, for example, did favor the simple-to-complex model, others evaluated on as regular a basis as did their colleagues

Table 4
Origin of Instruments
(Column Percentages Given in Parentheses)

	British Columbia		Ontario		Total
	Elem	Sec	Elem	Sec	
Borrowed From	93	46	67	33	239
Other Sources	(59)	(26)	(37)	(16)	(33)
Adapted From	10	26	35	57	128
Other Sources	(6)	(15)	(20)	(27)	(18)
From Own Files	5	6	20	18	49
	(3)	(3)	(11)	(9)	(7)
Created It Myself	49	98	57	101	305
	(31)	(56)	(32)	(48)	(42)
Total	157	176	179	209	721

in mathematics and biology, some of whom used the simple-to-complex scheme or some combination of both approaches.

Origination. Both provinces have provided evaluation resources for educators to use in their classroom assessments if they so choose. For 721 of their instruments teachers indicated where these had originated. (See Table 4 for a level-by-province breakdown of the sources of instruments.)

More of the secondary instruments than the elementary instruments were created wholly by the teachers. Fifty-two percent of the secondary instruments were reported as self-created while 32% of the elementary instruments were so described. Most of the rest of the instruments were either borrowed or adapted from other sources. In interviews with the teachers it was determined that these other sources might be published tests or item pools but were more often workbooks, texts, guidelines, and in British Columbia previously administered government examinations. The major complaint about outside evaluation tools was that they usually needed to be altered to fit the expectations of the particular class or course in which they might be used.

A total of 7% of the instruments, more in the Ontario sample than in the British Columbia sample, was reported to be located in test files. The rather insignificant use of previously used test materials may account for the almost total lack of response to an interview question concerning item tryouts, validation of keys, and item analysis generally. While one or two teachers reported using machine scoring for some tests, a scoring which produced difficulty level information for items, most had no experience with test statistics other than medians or means nor with the concepts of test or item analysis.

Item types. It was possible to categorize the different type of items used on the instruments in three ways. Certain items required the student to select information from choices given (alternative response, multiple choice, and matching), while others required the student to provide an answer (comple-

tion, short answer, essay, and construction). A third type required the student to perform a task, and that performance would be rated subjectively (anecdotal report) or objectively (checklist or rating scale) by the teacher.

Teachers typically reported using three or four of these item types in their evaluation, and defended their choices of types in a variety of ways. Some indicated that the age of the student meant that certain ways of evaluating were more appropriate. In both provinces, there was a general trend toward using performance measures at the primary level, while moving toward completion and, to some degree, multiple choice at the junior/intermediate grades.

At the secondary level in both provinces, the content of the subject was often given as a reason for the use of particular item types. Essay answers, for example, were cited as fitting with the objectives of particular courses better than other item types. But the pragmatics of evaluation also featured quite prominently in secondary teachers' reasons for item use. Conditions such as appearance on government examinations, need to diversify ways of collecting information, and ease of construction and marking were also frequently cited as reasons for use of particular item types.

The number of instruments reporting at least one example of each item type is recorded in Table 5. While much of these data could be shaped by particular teachers using many evaluations, there is a clear indication that short answers requiring single word or sentence responses characterize much of the evaluation at both levels. Objective performance items (checklists and rating scales) were also commonly used, especially in Ontario. The anomaly of the use of matching items at the secondary level in British Columbia can be explained by its appearance on some government examinations. It is the most common selection type of item in widespread use among the teachers in this sample although multiple choice was the third-ranked item in use by elementary teachers in Ontario.

Characteristics of the Judgments

For each instrument included, teachers were asked to rate the obtained achievement of their students compared to what the teachers' expected them to achieve. A consistent pattern emerged across levels and provinces concerning this aspect of judgment as shown in Table 6. More than half the time, the students performed about as well as teachers expected. For the total group about one fifth of the results were better than expected. This finding was reinforced by spontaneous comments by teachers about the relationships they expect between the measures they create and their judgments about achievement. Typical comments were as follows:

An elementary teacher in British Columbia, "Things do not change too much. A C+ in September is usually a C+ in June."

A secondary teacher in British Columbia, "Basically [you get] the same results if you give two tests or a thousand, it ranks them the same way."

An elementary teacher in Ontario, "Each year the tests are tailored to the class."

Table 5
Number of Instruments Using Particular Item Types by Level and by Province
(Percentages of Total Given Within Columns)

Item Type	British Columbia		Ontario		Total
	Elem	Sec	Elem	Sec	
<i>Selection</i>					
Multiple	5	19	38	19	81
Choice	(3)	(7)	(18)	(6)	(8)
Matching	0	34	5	8	47
	(0)	(13)	(2)	(2)	(5)
Alternative	6	14	9	7	36
Response	(3)	(5)	(4)	(2)	(4)
<i>Supply</i>					
Completion	87	81	76	59	303
	(47)	(30)	(36)	(18)	(30)
Short Answer	8	61	10	60	139
	(4)	(23)	(5)	(18)	(14)
Essay	8	16	9	50	83
	(4)	(6)	(4)	(15)	(8)
Construction	9	18	18	45	90
	(5)	(7)	(8)	(14)	(9)
<i>Performance</i>					
Objective	22	26	46	65	159
Performance	(12)	(10)	(21)	(20)	(16)
Subjective	42	2	3	14	61
Performance	(22)	(1)	(1)	(4)	(6)
Total	187	271	214	327	999

Table 6
Teachers' Expectations of Student Performance
(Column Percentages Given in Parentheses)

EXPECTATIONS	British Columbia		Ontario		Total
	Elem	Sec	Elem	Sec	
Better than expected	27	46	59	38	170
	(15)	(26)	(32)	(16)	(22)
About as well as expected	79	96	104	137	416
	(45)	(54)	(57)	(58)	(54)
Poorer than expected	11	26	5	33	75
	(6)	(15)	(3)	(14)	(10)
Other or omit	60	10	16	29	115
	(34)	(6)	(9)	(12)	(15)
Total	177	178	184	237	776

A secondary teacher in Ontario, “When you’ve been teaching that long, you’ve got a pretty good handle on it. By the time I get the names down, I have a pretty good idea of where they are.”

Such expectations indicate that the teachers predict a pattern of results based on their previous experiences with the students forming a particular class. However, they contradicted the analysis in Table 6 when, in describing their instruments on the checklist, they only checked 20% of them as confirming an informal assessment.

Characteristics of the Decision Making

Each teacher was asked to rate the purposes of each instrument with no restriction on the number of purposes each could serve. The results are given in Table 7. The overwhelming purpose checked for most instruments was the generation of marks for reporting purposes (81%). At the elementary level in both provinces, the need to diagnose students’ weaknesses with material was also a frequently cited purpose. At the secondary level in both provinces, the checking of students’ progress against course objectives was the second most frequently checked purpose.

Certain purposes can be deemed important to the teaching-learning process while others are important for external purposes. At the elementary

Table 7
Percentage of Instruments Citing Purposes of Evaluation
(Rank Within Column Given in Parentheses)

PURPOSES	British Columbia		Ontario		Total
	Elem	Sec	Elem	Sec	
Marks	86 (1)	80 (1)	77 (3)	81 (1)	81
Criterion-Reference	44 (4)	70 (2)	71 (4)	78 (2)	67
Apply learning	48 (3)	60 (4)	82 (2)	67 (3)	65
Diagnosis	56 (2)	47 (8)	90 (1)	37 (8)	56
Self-Monitoring	28 (8)	53 (5)	62 (5)	41 (7)	46
Prepare for future exams	28 (9)	62 (3)	29 (8)	48 (5)	42
Assignment complete	26 (10)	50 (6)	29 (9)	49 (4)	39
Self-Teaching	35 (7)	47 (7)	43 (6)	33 (9)	39
Norm-Reference	25 (11)	30 (9)	25(11)	42 (6)	31
Teach next	40 (5)	23 (10)	37 (7)	5 (11)	25
Other	37 (6)	5 (11)	28(10)	13 (10)	10

Note: *Marks*: to generate marks for reporting; *Criterion-Reference*: to check students’ progress against course objectives; *Apply learning*: to have students practice or apply what has been learned; *Diagnosis*: to diagnose students’ weaknesses with the material; *Self-Monitoring*: to enable students to monitor their own progress; *Prepare for future exams*: to prepare students for this kind of evaluation in the future; *Assignment complete*: to ensure that students do assigned work; *Self-Teaching*: to allow me to see how well I taught the material; *Norm-Reference*: to compare students’ achievement to others; *Teach next*: to help me decide what to teach next.

level in both provinces, the five top-rated purposes were concerned with informing the teaching-learning process except for the generation of marks for reporting purposes. At the secondary level in both provinces, a concern with preparing for evaluation in the future along with mark generation were important aspects of purpose. For the elementary teachers generally, the lowest rated purposes were those external to the teaching-learning process: to compare students' achievement to others, to ensure students do assigned work, and to prepare for future evaluations. The lowest ranked purposes for secondary teachers in both provinces were concerned with diagnosis and reteaching: deciding what to teach next, diagnosing students' weaknesses with the material, and allowing teachers to see how well the material had been taught.

During the interview, teachers were asked to rate their evaluation in terms of criterion-referenced and norm-referenced purposes, both in terms of their own perceptions and what they thought their students would perceive. In the primary classrooms, the issue of norm-referencing does not generally arise. As the grade level in the sample increased, however, so did the report that the students would perceive the evaluations in a normative or in some combination of normative and criterion ways. At the secondary levels in both provinces, the teachers were inclined to see the instruments they developed as objectives-based, a view that is supported by the frequency with which the checking of the students' progress against course objectives purpose was noted. These teachers recognized, however, that in some respects their students were right in seeing their results in a comparative way, at least in some limited respects.

A final decision making issue concerned itself with what teachers and students were expected to do with the results of the evaluation process. Almost all the teachers in the sample reported a process of "taking up" or "going over" the completed instruments in class. At the elementary level, this process was frequently supplemented by encouraging the students to take the corrected instruments home. At the secondary level, students were more often encouraged to file them for use in final examination preparation. Several teachers at this level reported collecting the instruments so that they could be adapted for successive classes, and six secondary teachers reported that they had a retesting policy for students who had done poorly.

As for the teachers' own use of the results, the elementary sample often gave specific alterations in methods and content they would use the next time they taught that particular unit although the frequency of this type of comment decreased as the grade level increased. At the secondary level, particularly in British Columbia, there was little to no change expected in methods and content, although a few teachers mentioned that their expectations for students might be altered. In Ontario at the secondary level, the teachers were more likely to make general comments about the pacing of the instruction, for example, than was the case in British Columbia, but still the overall sense was that the content and method were constant and stable.

Discussion

The teachers' concerns with measurement characteristics of the instruments they use focus on content validity and on centering the items to the students' abilities. Elements that shape much professional evaluation practice—test blueprinting, item analysis, statistical manipulation of data—are virtually absent from classroom practice. Indeed, the search for objective measurement is deliberately offset by the frequently-expressed desire to match the measurement results to expected results. Items and tasks tend to be required that students are likely to be successful in, that look like tasks that were done in the regular flow of classroom work, and that support conclusions about individual and class achievement already made on other bases, including the results of other instruments. Such an approach inhibits the use of instruments prepared by others unless these can be adapted to specific groups of students.

The judgment that flows as a result of this activity, then, is largely one of confirmation. With more objective measurements, discrepancy analysis could be used. But when the object of the measurement activity is to produce information that is congruent with what was previously known, then the possibility of discrepancies being noted is limited. The cases of teachers preparing students for government examinations are enlightening in this regard.

Where the provincial examination mark is to be combined with classroom-generated marks, the expectation might plausibly be made that the classroom activities would concentrate on those objectives and activities that cannot reasonably be evaluated in paper-and-pen fashion. What actually happens in our sample of teachers is a tendency to imitate the provincial examination in virtually every instrument. Not only does this give the students practice in the types of questions featured in such examinations, but it allows the teachers' judgments of the students to correlate highly with those emanating from the single, end-of-year assessment.

A notable exception to this general, confirmatory pattern occurs at the primary and early primary levels. Here, the teachers were more likely to stress as major purposes of evaluation those that dealt with the teaching-learning process. They were also more likely than their later elementary or secondary colleagues to report using the information generated by the instruments to alter their teaching activities, particularly for individual students.

It is also the case that this particular group of teachers operates in a policy and procedural framework that is quite different from that of their colleagues. Report cards are more likely to be anecdotal and personal and less numerical and comparative. Outside examinations are infrequent or nonexistent. Preparing for future schooling or work where school marks may be used as criteria is not relevant. Consequently, the purposes these teachers had for evaluation rated those appropriate for internal classroom purposes much more highly than those purposes were rated by secondary teachers.

Where the outcomes of the assessments serve primarily to confirm judgments already made about relative achievement, there will be little that teachers can do to use the information for either self-improvement or in-

dividual student diagnosis. That this confirming purpose overrides these other potential uses is verified by the use made of the instruments' results. They tend to be analyzed in terms of the rationale for the judgment, and then stored, to be used later as contributor to single, numerical summaries of individual achievement.

Decision making, in the sense of making choices among courses of action, is not a plausible expectation when the dominant aim of measurement is to confirm. The rational decision making model depends on the possibility of divergent outcomes, with each outcome associated with a different course of action. The most sensible course of action to proceed with in the case of confirming information is to carry on as usual—which is precisely what most of the teachers in the study reported doing.

Conclusion

The evaluation of student achievement by teachers in classrooms occurs within a policy and procedural framework that is largely determined by outsiders. As other parts of our study are in the process of demonstrating, the department and school, the district or board, and the provincial authorities all weigh in with regulations and expectations that color much of the data-gathering activities teachers pursue.

Where that does not occur so much, at the primary levels, the teachers tend to use their measurement and evaluation devices to comment on the growth of individuals and the success with which a particular unit or activity worked. Beyond those grade levels, however, the procedures for reporting especially mean that most teachers proceed with their evaluations primarily because they must produce marks, and these marks must be in a prescribed, usually norm-referenced, form. At the tertiary end of public schooling, at least in British Columbia, there is a further check on the congruence of the teachers' assessments in certain subjects through a formal examination program. One certain effect of this program is to bring the teacher-generated marks in line with the examination outcomes.

From early on in the schooling process, the formative purposes of evaluation begin to give way to summative ones, from internal classroom uses to external, public ones. One deleterious effect of this movement—if the teachers involved in this study are representative of their colleagues and their systems—is to encourage instructionally irrelevant evaluation, of replacing what might enhance the teaching-learning process with a mark-driven ritual.

Acknowledgment

The research reported here was supported by a grant from the Social Sciences and Research Council whose help is gratefully acknowledged.

References

- Frisbie, D.A., & Friedman, S.J. (1987). Test standards: Some implications for the measurement curriculum. *Educational Measurement: Issues and Practice*, 6(3), 17-23.
- Gullickson, A.R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research*, 77(4), 244-248.
- Mayo, S.T. (1964). What experts think teachers ought to know about educational measurement. *Journal of Educational Measurement*, 1(1) 79-86.

- Rees, R. (1989). *The influence of the provincial educational system on educators' evaluation policies and practices*. Paper presented at Ontario Educational Research Council Annual conference, Toronto.
- Stiggins, R.J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22,(4), 271-286.
- Wilson, R.J. (1989). Evaluating student achievement in an Ontario high school. *Alberta Journal of Educational Research*, 35(2), 134-144.
- Wilson, R.J., & Rees, R. (in press). The ecology of evaluation: Measurement and evaluation within an administrative framework. *Canadian Journal of Education*.

PHILIP NAGY

Ontario Institute for Studies in Education

ROSLYN MOORHEAD

The University of Western Ontario

Administrative Response to Classroom Testing Data: A Problem Solving Perspective

A group of 15 practising high school principals were presented with a case study to examine their responses to classroom testing and the problems associated with it. The case contained information on a district-wide report outlining inconsistencies in teachers' evaluation practices, with a parental complaint concerning the evaluation practices of a staff member. Response protocols were categorized two different ways: first, into global areas of concern—student, teacher, parental complaint, district report, school department, school, district, and problem solving method; and second, by type—reasoning, action, reiteration, values, background, summary, and comment. There was little evidence that the statement types and patterns of explanation bore a relationship to the quality of actions individuals might take. The study may have been tapping differences in think-aloud ability rather than administrative ability. However, only about half the subjects related the district-wide study to the parental complaint. The study also examined the methodology used. Although the analysis began from a general skills perspective, context-specific elements had to be added to complete the task. Implications for the analysis of problem solving data are explored.

Introduction

The intent of this report is to set the issue of classroom testing and problems associated with it in the broader context of the school program and its ongoing maintenance. To accomplish this end, we presented practicing high school principals with a case study and examined their responses to the situation presented. The case contained information on a district-wide report outlining inconsistencies in teachers' evaluation practices, as well as an immediate parent complaint about the evaluation practices of a staff member.

In addition to dealing with the substance of the principals' reactions, we also provide some detail on our methods, from cognitive science, of examining the transcripts provided by our sample. Apart from questions involving this specific data set, we hope that our methodological comments are relevant to a broader issue in measurement, that of the systematic analysis of complex student achievement data. In particular, and very much in keeping

Philip Nagy is an associate professor in the Department of Measurement, Evaluation, and Computer Applications. His interests are in applied measurement and the assessment of problem solving.

Roslyn Moorhead is a research officer in the Educational Leadership Centre. Her interests are in school administration and principal development.

with a recent article by Perkins and Salomon (1989), we began our analysis from a general skills perspective and found ourselves needing to add context-specific elements to complete our task. The reasons for these modifications shed light on the analysis of problem solving data, whether the subjects be students, teachers, or principals.

Background

The rationale for this line of inquiry rests on three points: first, inconsistencies in classroom testing practice are a serious problem; second, the principal is a key figure in contributing to the solution of this problem; and third, our analysis methods, from the literature on problem solving, are appropriate for the task we have set ourselves.

Rationale

Testing Practice

Measurement specialists have begun to make teachers' assessment practices a target for examination. In an example of this line of inquiry, Traub, Nagy, MacRury, and Klaiman (1988) examined the assessment practices of high school calculus teachers, all from large, relatively wealthy school districts in the urban core of Southern Ontario. The teachers kept daily logs of their teaching and evaluation activities and provided copies of their tests, quizzes, and exams for an entire semester. In addition, they all marked the same set of student exams obtained from a school outside the study and provided by the research team. The results of the investigation showed considerable variation in teacher practice. Teachers differed substantially in use of classroom time, in the amount of student effort required, in relative emphasis given topics, in match between content taught and content tested, in methods used for assessing student achievement, and in overall generosity in awarding marks.

As well, Traub et al. (1988) have provided a convenient, although brief, review of related work. They report:

1. Teacher-developed tests and classroom observations are the most common methods of assessing students (Herman & Dorr-Bremme, 1983);
2. Variations in assessment practices exist across subject areas and grades (Green & Stager, 1986; Terwilliger, 1987). For example, elementary teachers tend to favor observations, and secondary teachers testing (Stiggins & Bridgeford, 1985); and mathematics is heavily test-oriented compared with other subject areas (Wahlstrom & Danley, 1976);
3. Teachers need training in teaching and assessing higher-order thinking skills (Haertel, 1986; Stiggins, 1988); and
4. The measurement community needs to learn more about how assessment is conducted in the classroom by using more focused research methods such as classroom observation (Stiggins & Bridgeford, 1985; Stiggins, Conklin, & Bridgeford, 1986; Anderson, 1987).

The Traub et al. (1988) results were incorporated in a case study presented to a sample of principals in this study. The case study (Appendix A) contains a summary of the results obtained by Traub et al., modified only by

changing the number of schools involved from 17 to 15 and by placing these schools in one school district rather than 13.

The Role of the Principal

Current administrative theory assigns the task of curriculum leadership and management, including student evaluation practices, to the principal. Rutherford (1985) has outlined concrete aspects of effective school leaders. Among his indicators of effectiveness are the setting of clear expectations for both teachers and students, knowledge of individual classrooms, instructional support, and monitoring of progress. Leithwood and Montgomery (1986) identify four levels of principal effectiveness which focus, from lowest to highest, on the rules, the climate, the program, and the students. Among the factors under the principal's rather than the classroom teacher's control, they include assessment, recording, and reporting procedures. At the policy level, the expectation of the Ontario Ministry of Education (1988), stated in their recent manual on curriculum management, is that the principal will assume responsibility for achieving the goals of the curriculum. From both the theoretical and policy perspectives, a major role of the principal involves monitoring of the degree to which the objectives of the curriculum have been accomplished, that is, student evaluation and assisting teachers in the improvement of their practice.

Problem Solving

Our strategy for analyzing the data collected in this study is rooted in the artificial intelligence approach to how people solve problems. Problems can be placed on a continuum from well-structured (logic puzzles, chess, and physics problems) to ill-structured (economic, social, and political problems). Well-structured problems are characterized by low linguistic load, little reliance on context, and widely accepted procedures and correct answers. Ill-structured problems are characterized by high linguistic load, much reliance on context, and debate about accepted procedures and solutions. Our focus, principals' responses to case studies, is an example of an ill-structured problem.

The most immediate source for our analytic methods is the work of Voss and his colleagues at The Learning Research and Development Center, University of Pittsburgh (Penner & Voss, 1983; Voss, Greene, Post, & Penner, 1983). They analyzed responses to protocols on the question of (ill-structured) Soviet agricultural production problems, using parallel structures for problem solving and verbal reasoning. Examples of steps ("operators") in the problem solving structure include "state subproblem," "provide support," and "summarize." In the verbal reasoning structure, they categorized subjects' responses as, for example, "state assertion," "elaborate," and "state qualifier."

Earlier work (Nagy & Allison, 1988), involving responses from a small sample of experienced principals to a different case study, provided an initial adaptation and refinement of the Voss et al. (1983) methods. In that analysis, Nagy and Allison first categorized statements from protocols into a set of 19 categories (e.g., reasoning, subproblem, summary). They then drew diagrams

in which statements were grouped by subproblem; a system of indentations showed subordinate relations between statements. For example, a reasoning statement was indented once from the action it justified, and a subsequent elaboration of that reasoning was indented a second time. Finally, they drew highly simplified summary diagrams showing the pattern of subproblems identified and actions taken.

In the present analysis, we began with essentially the same system. We found, however, that the detailed statement categorization caused us to dwell more on the language of the respondent than on the substance of his or her response to the case and that our category system failed to provide a basis for judgments of quality (e.g., for differentiating poor problem solving from better). Thus we greatly simplified the statement categorization system and paid more attention to the subproblems addressed by the principals and the specific actions that would have been taken. In retrospect, we question the extent to which methods involving categorization of statement types tap verbalization ability compared with problem solving ability. On this issue, we have some suggestions for refinement of the methodology.

Method

Sample

The principals who participated in this study were chosen for convenience—essentially from personal knowledge and for geographical location. As about half came to our attention through supervisor or peer recommendation, the sample should probably be considered slightly more skilled than a random sample of experienced, practicing principals. However, apart from knowing that three of the 15 are in their first or second year at the principalship, we have no systematic way of distinguishing relative expertise. All interviews were conducted in the fall of 1988, in the offices of the principals, and at their convenience.

The Case Study

The hypothetical case study (Appendix A) contains two elements. The first element involves results of a research study recently completed in the high schools of the district. The study, using one calculus teacher from each school, showed large differences (a) in use of class time (e.g., seatwork versus direct teaching), (b) in amount of practice given, (c) in content covered, (d) in amount of evidence collected for grade assignment, (e) in match between content taught and content tested, and (e) in grades assigned to a common set of student exams. The study was reported to the board, where it occasioned some comment. The principal in the situation had previously asked his math head and department to discuss it, and they reported no problems in their school. The second element of the case study is a parental complaint. The student in question, reported by her father as having done well in math in the past, is depicted as having difficulties in calculus. The father wants a change from one teacher to another. In the scenario, the responding principal is asked to place himself or herself in the shoes of the principal in the case study and describe what he or she would do in the situation. The interviewer

answered clarifying questions, but no probes were used to explore areas that did not spontaneously occur to the responding principal.

In order to understand the context, the reader might find it helpful to know that calculus is taught in Ontario high schools only in the final, the Ontario Academic Credit (OAC) year. For most students, this is the fifth year of high school. However, the system is designed so that the more able, if they are willing to take a “no-frills” curriculum, can finish high school with university entrance requirements complete in four years. For further context, there are no provincially set exams nor any equivalent of the Scholastic Aptitude Test in Ontario. High school grades are the major basis for university entrance, and they are determined entirely by the teacher.

The Analysis

In the initial analysis, response protocols were segmented into thought units and numbered sequentially. A “thought unit,” intuitively defined, was essentially a sentence; compound sentences were split, and items listed in a sentence were individually numbered. Next, these units were categorized into about 20 categories using a system adapted from Nagy and Allison (1988). Finally, diagrams were drawn to show the flow of the problem solution. In these diagrams, flowchart-like symbols for statement types were used, sequence was shown by vertical (downward) shifts, and change in subproblem area by left-to-right shifts.

As described above, this approach was judged unsatisfactory because it dealt too much with the articulation of the problem solving and not enough with the substance. We felt that we were dissecting differences in oral communication ability rather than problem solving ability. More importantly, it did not provide a systematic method of describing what each individual chose to do or not do in the presented situation. It left the distinction between conscious linking of subproblems and uncritical meandering between subproblems to the unaided subjective judgement of the analyst.

The revised analysis proceeded from two perspectives: that of the problem solving processes used, and that of the issues addressed. Two major changes from the first attempt were undertaken. First, the preliminary analysis had given us a sense of the range of areas addressed and actions taken. We made a decision to assign each statement uniquely to one of eight areas. (This required some resegmentation into much smaller thought units.) Second, the number of statement types was greatly reduced, from about 20 to just eight. Mainly, a large variety of subtle distinctions were ignored in favor of one larger category, “reasoning.”

The analysis reported has four components: first, the emphasis given to each area; second, what was actually accomplished by each respondent; third, the statement types used; and fourth, the explicit value statements made by the respondents. Where possible, we attempt to relate these different views of the data to each other.

Results

The responses were segmented into statements. With one low outlier (Principal #1, 23 segments) and two high outliers (#4, 461 segments, and #6, 259

segments), the responses clustered quite closely around 100 segments. Preliminary analysis showed that the respondents dealt with eight different areas:

- 1. *Student*—gathering information on the student, talking with him or her, or setting up assistance.
- 2. *Teacher*—talking to the teacher, or getting information on the teacher’s performance (also includes any dealing with the receiving teacher).
- 3. *Parent/Complaint*—talking to the parent, trying to resolve the issue; excludes any information gathering needed for resolution of the conflict.
- 4. *Study*—discussion of the study of assessment practices in the abstract; that is, not in the context of what the department, the school, or the district ought to do.
- 5. *Department*—talking to the four teachers of calculus as a group, rather than the teacher whose student has complained; or talking to the department head; discussion of procedures the department ought to have in place, or information that ought to be available.
- 6. *School*—discussion of school-wide issues related to the case; response had to involve explicitly other departments beyond the math department, or clearly talk of “departments” or “heads” in the plural.
- 7. *District*—discussion of issues involving the elected board, the district central administration, or district level committees.
- 8. *Methods*—any discussion of how to solve problems, or how to deal with conflict.

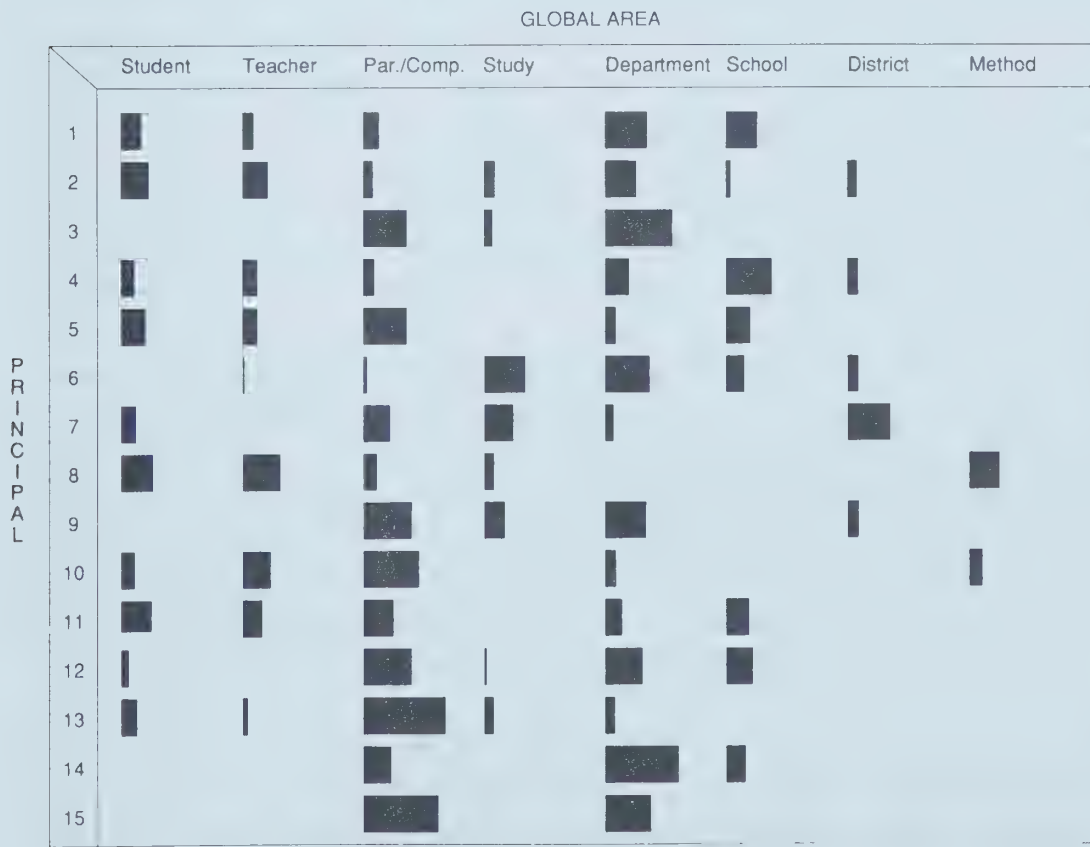


Figure 1: Percentage statements by global area

Figure 1 shows the degree of emphasis in each principal's protocol on the eight areas. The sum of the widths of the boxes in Figure 1 for each principal is a constant, representing 100% of the statements given in response to the case. As can be seen, while all respondents paid some attention to the parent making the complaint, only two thirds of them talked directly to the student or the teacher. Only one third dealt with issues beyond the school boundaries, and of these (not shown in the figure), only one made a reference beyond the school district (i.e., to the effect that students had to compete with students from other schools for admission to university). Just over one half of the respondents discussed issues raised in the research study apart from the specific parental complaint, while only two dealt in a meta-analytic fashion with problems or conflict. All but one of the respondents discussed departmental procedures. A few of the respondents (#3, #14, #15) can be seen to have addressed only a limited number of the eight areas, but most dealt with five of the eight.

Table 1 gives a matrix of what was accomplished or addressed by each principal. These accomplishments do not have a one-to-one correspondence with the actions taken; several actions might be undertaken for one accomplishment, for example, to find information on the student. As well, there are only seven categories in Table 1, compared with eight in Figure 1; no aspect of the discussion of the study could be considered an accomplishment apart from accomplishments involving the teacher, the department, or the school. An examination of Table 1 shows that only six of the 15 responding principals would have checked to see if the information from the parent on the student was accurate. However, eight of 15 set up some type of assistance, either academic or personal, for him or her. Except for two respondents, there was generally less investigation of the individual teacher than of the student. Only about one half the sample tried to find out what the department did to monitor or control teacher practices with respect to the topics examined in the district study. In effect, such investigation shows a clear linkage between the individual complaint and the general issue. Thus one half the sample treated the study and the complaint as only marginally related.

The information in Table 1 sheds light on how the principals felt about the issues raised in the reported study of calculus teacher practices. Three respondents talked of dealing with the specific complaint within the school. Two of these individuals were concerned to avoid the imposition of a solution (to the general issue rather than the complaint) from outside the school, while the third talked of the political ramifications of having a hostile parent go to the district administration. Only three respondents stated that the district office needed to be involved in resolving between-teacher inconsistencies.

To further explore how respondents felt about the issues, we isolated their value statements (see Appendix B). About half stated that the promotion of uniformity of marks within the school was important, but only two spoke specifically in favor of common exams, even within the school. Most talked of ongoing communication among teachers as the main vehicle for the promo-

Table 1
Matrix of Activities

	Principal														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>Student</i>															
academic info	1			4	5			8			11		13		
personal info	1			4	5						11	12	13		
academic assistance							7	8		10					
personal assistance				4	5					10	11		13		15
direct contact	1	2									11	12	13		15
learning style							7					12		14	
<i>Teacher</i>															
competence info	1							8							
attitude info		2			5			8				12			15
study-like info		2						8		10	11		13		15
academic assistance				4				8		10					
personal assistance				4									13		
flexibility info				4						10					
other teacher info										10					
<i>Department</i>															
study-like data			3	4					9	10	11		13		15
promote conformity	1	2	3	4					9			12		14	
don't promote						6									
use one teacher							7					12			
flexibility info										10					
tutorial set-up	1														
success info						6									
dismiss report												12			
<i>School</i>															
transfer info	1				5										
department info															15
contain problem					4	5				10					
<i>Parent</i>															
meet with principal	1	2	3		5								13	14	15
meet with teacher				4		6	7		9						
discourage move			3	4					9		11				
assess if vocal				4											
ongoing talk										10					
dismiss complaint														14	
<i>District</i>															
admin view—study		2													
avoid system exams				4											
ignore trustee						6									
system committee							7								
system policy				4											
<i>Other</i>															
general strategy								8							

tion of uniform marking standards. Also, in only a couple of cases is it explicit that the motivation for achieving consistency is fairness for students rather than the avoidance of complaints.

As a final step in our analysis, we categorized statements according to type:

Reasoning—a general term including several categories of the original analysis: identifying subproblems, pointing out assumptions, citing information from the case, identifying missing information, predicting, explaining an action, analyzing the case; *Details* of Reasoning statements are tabulated separately.

Action—specific statements of what the respondent would do; again, *Details* of Action statements are tabulated separately.

Reiteration—simple repetition, which was distinguished from attempts to provide a summary.

Values—statements of what the respondent considers important about assessment, students, or education; where detail or elaboration was provided, it was not tabulated separately.

Background—relating the case to the respondent's personal experience; in two situations, a respondent either began or ended the discussion with a lengthy personal anecdote; these were simply not analyzed.

Summary—summaries of previous steps.

Comment—statements about the process itself, including both throwaways ("Where was I?") and substantial meta-analytic statements ("I like to lay out the steps in a problem solving process before I begin.").

Figure 2 gives a breakdown of the statement types used by each respondent. Again, the sum of the width of boxes for each respondent is a constant, representing 100% of the statements. For Reasoning and Action, the bar down each box distinguishes Reasoning or Action, on the left, from Detail, on the right. In the two cases with no bar, there was no detail provided. One interesting comparison to make in Figure 2 is between Reasoning and Action. For about one half the sample, there were more than twice as many Reasoning statements as Action statements.

There are several other patterns within Figure 2, but their relationship to the quality or nature of response to the case is not clear. Voss and his colleagues (Penner & Voss, 1983; Voss et al., 1983) suggest in their research that at least two factors differentiate the expert from the novice: experts isolate more abstract or general problems as being the primary factors producing a given problem, and experts show more skill in providing supportive reasoning for their stated solutions, in stating subproblems which may emerge from the solutions, and in evaluating solutions in relation to constraints. Neither of these patterns is apparent here, however.

Subproblem identification in our protocols was frequently implicit rather than explicit. None of the respondents consistently articulated or defined subproblems associated with the case; rather, subproblems were implicitly identified in the stated solutions or action plans. Hence a category for subproblem identification has not been included as one of the seven statement types. Instead our analysis focused on the matrix of activities and the areas

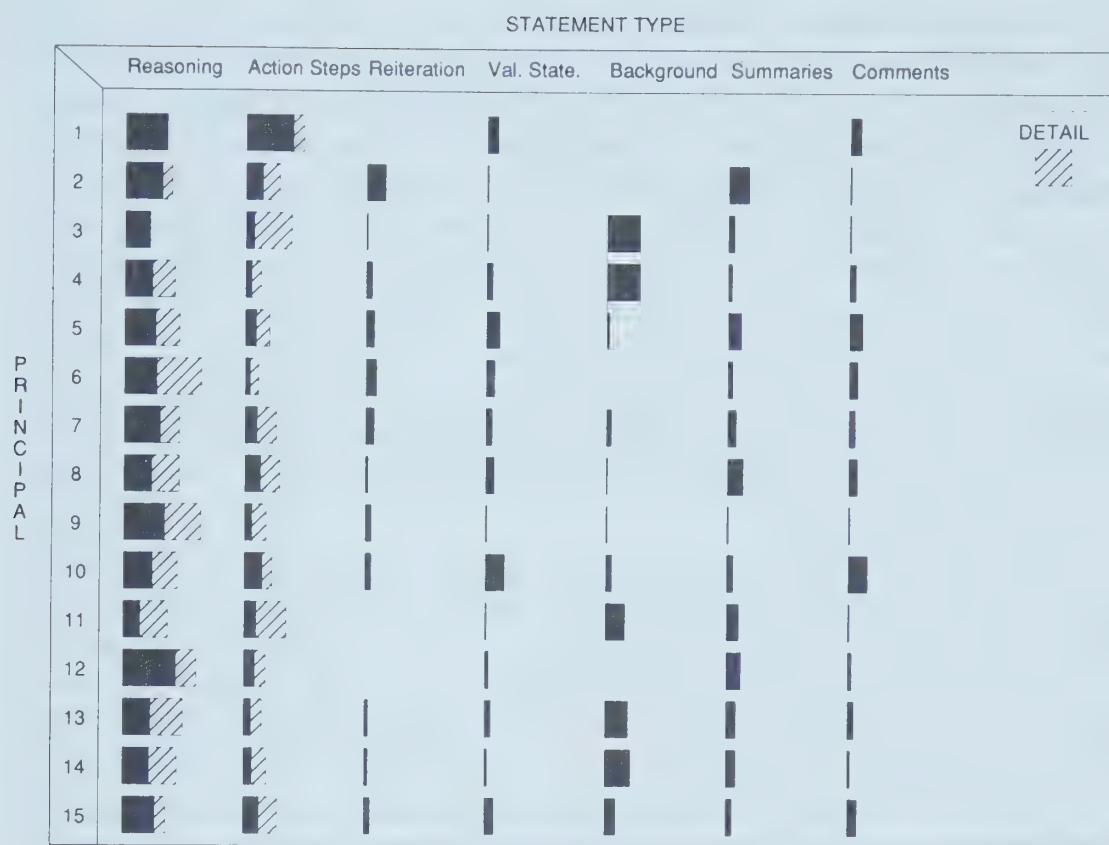


Figure 2: Percentage statements by type

commented on by the respondents. We found that everyone dealt with the immediate problem of responding to the parent, but only half dealt with the broader or more abstract issues concerning the study or the ramifications for the school, and only one third mentioned the school district. However, of the three novice principals (those with less than three years experience), one (#2) talked of all three areas, and another (#4) talked of two of them.

In terms of reasoned arguments, some respondents reiterated and backtracked more than others; some provided periodic summaries; others related their actions strongly to their own school situation; and some expressed many more values than others. Again, however, there was no correlation between the extent of these activities and the experience of the principal. In fact, a novice (#4) was one of those who employed substantial supportive reasoning. Moreover, in our subjective opinion, more reasoning is not necessarily better; some respondents engaged in critical analysis of the problem and others in uncritical meandering. The present analysis by statement types does not differentiate between these two.

Discussion

Several issues, both methodological and substantive, arise from this investigation. With a reminder that we believe the analysis of complex problem solving data should be of interest to the measurement community, we begin with comments on our own methodology.

First, there is little evidence in the present data set to suggest that statement types used and patterns of explanation bear a relationship to the

quality of action the individuals might take when placed in a situation similar to the presented case. We are at this point unable to relate differences in ability to articulate one's views to ability to run a school successfully. Similarly, there is little reason to believe that patterns of statement use, such as in Figure 2, can be related to substantive issues involving student assessment attitudes and practice. One explanation, which we are currently pursuing, is that we have a restricted-range problem. Research on expert-novice differences (e.g., Patel & Groen, 1986) has shown differences in expert use of "forward reasoning" and novice use of "backward reasoning." By such a criterion, all our sample were experts. We may indeed find the data on statement type useful in differentiating between responses from practicing principals and true novices such as student teachers, or less experienced classroom teachers.

In addition, we felt the need to introduce Table 1, which amounts to a subject specific "marking scheme," the subject being school administration. This is in keeping with Glaser's (1984) argument for the importance of specific knowledge over general skills. Using a content-based system, we can distinguish levels of success at the task. However, as Perkins and Salomon (1989) point out, more atypical problems might better allow differentiation between experts and novices in general skill (i.e., the approach in Figure 2) rather than subject-specific skill (i.e., the approach in Table 1). Perhaps our case study was not novel enough for our sample. At the time, we found comments such as "this is a real problem—it happened last week" quite gratifying, showing that we had not lost touch with school realities. However, a truly novel problem may have allowed differentiation in terms of general skills as well as subject-specific accomplishments.

Third, we felt discomfort that we were tapping differences in verbalization skills rather than school administration skills. This suggests development of a substantial set of probes to compensate for verbalization ability. With such probes, we would still be able to distinguish what is spontaneously offered by respondents, but we would be less in the dark about all the things not said. That is, we would know if a given respondent had actively chosen not to take some action or had merely forgotten to mention it in the discussion. As Nagy and Allison (1988) pointed out, our experts are experts at doing things, rather than, as were Voss' experts, at talking about things. However, this argument cannot be carried too far; one might reasonably argue that principals earn their living with words as much as do academic agricultural experts.

On the substantive side, we found reasonably consistent desires among our sample for grading consistency, but large differences in how large a role to play in achieving this consistency. Only a minority took an active hand in departmental discussions; most left it to the department chair. This view is not in keeping with the view of the principal as curriculum leader. However, it does seem to represent current practice in Ontario high schools.

Second, and this may be a methodological point poorly disguised, we have tapped the "craft" of administration. Our subjective impression is that our group of respondents, all successful by different criteria, actually differ less

in administrative ability than our results show. This suggests that we need to explore the importance of rationality—the requirement for success at our task—to actually running a school.

Finally, and perhaps most important for classroom assessment, only about one half our sample actually treated the study described in the case study as relevant to the parent complaint. This is perhaps the most important food for thought for those involved in the improvement of teachers' assessment skills.

References

- Anderson, J.O. (1987, June). *Teacher practices in and attitudes towards student assessment*. Paper presented at the Annual Meeting of the Canadian Educational Researchers' Association, McMaster University, Hamilton, ON.
- Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, 39, 93-104.
- Green, K.E., & Stager, S.F. (1986). *Effects of training, grade level, and subject taught on the types of tests and test items used by teachers*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.
- Haertel, E. (1986). *Choosing and using classroom tests: Teachers' perspectives on assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
- Herman, J.L., & Dorr-Bremme, D.W. (1983). Uses of testing in the schools: A national profile. *New Directions for Testing and Measurement*, 19, 7-17.
- Leithwood, K.A., & Montgomery, D.J. (1986). *Improving principal effectiveness: The principal profile*. Toronto: OISE Press.
- Nagy, P., & Allison, P. (1988). *School-level problem solving: A cognitive perspective*. Paper presented at the Annual Meeting of the Canadian Society for the Study of Education, Windsor, ON.
- Ontario Ministry of Education. (1988). *Curriculum management*. Toronto: Ontario Government Publications.
- Patel, V.L., & Groen, G.J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science*, 10, 91-116.
- Penner, B.C., & Voss, J.F. (1983). *Problem solving skills in the social sciences: Methodological considerations*. University of Pittsburgh: Learning Research and Development Center.
- Perkins, D.N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, 18(1), 16-25.
- Rutherford, W.L. (1985). School principals as effective leaders. *Phi Delta Kappan*, 67, 31-34.
- Stiggins, R.J. (1988). *The nature and quality of teacher-developed classroom assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.
- Stiggins, R.J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), 271-286.
- Stiggins, R.J., Conklin, N.F., & Bridgeford, N.J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice*, 5(2), 5-17.
- Terwilliger, J. S. (1987). *Classroom evaluation practices of secondary teachers in England and Minnesota*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC.
- Traub, R., Nagy, P., MacRury, K., & Klaiman, R. (1988). Teacher assessment practices in a senior high school mathematics course. *Final Report, Transfer Grant 52-1028*. Toronto: The Ontario Institute for Studies in Education.
- Voss, J.F., Greene, T.R., Post, T.A., & Penner, B.C. (1983). Problem solving skill in the social sciences. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research theory* (pp. 165-213). New York: Academic Press.
- Wahlstrom, M.W., & Danley, R.R. (1976). *Assessment of student achievement*. Toronto: Ministry of Education.

Appendix A

Student Evaluation Case Study

Background

Smithson Collegiate Institute (SCI) serves a predominantly affluent, professional suburban community; however, there is one feeder school in a lower middle class neighborhood with subsidized housing. There are just over 1,400 students, substantial numbers of whom go on to university after graduation. Mr. Lesley is a recently appointed principal to Smithson Collegiate, just beginning his second year.

A district research study on student evaluation in calculus classes has just been completed and a summary sent to each secondary school, along with a memo from the assistant superintendent of program requesting that schools review the results and their implications. A summary of the 80-page report is appended.

The Issue

When the results of the study were presented at a board meeting, there was considerable discussion. One influential and experienced trustee stated that he expected that the situation was even more serious among less experienced teachers and in more difficult to evaluate subject areas.

Mr. Lesley sent this information to the math department head, Mr. Addison, who has held that position for 11 years, the past six at Smithson Collegiate. There are four calculus classes in the school, with four different teachers. The calculus course coordinator is a senior teacher with an excellent reputation among his peers for the quality of his teaching. The other three have several years of experience and are regarded as being competent. Mr. Addison told Mr. Lesley that he was not aware of any problems such as those mentioned in the report. Mr. Lesley asked him to raise the issue at the next department meeting and to keep him informed. The head agreed.

All department meeting minutes are routinely sent to Mr. Lesley, and the next set for the math department meeting included an item on the calculus study. Mr. Lesley noted that after some discussion, the teachers involved felt that the problems reported were not significant at SCI because there was a common calculus exam and a policy of no exemptions. The department head appended a note saying that, although there were differences of opinion regarding teaching emphases, he felt strongly that the calculus classes were well taught by experienced staff members, each of whom had individual strengths, and that any attempts to impose greater uniformity would be counterproductive.

A week later, an angry parent demanded to see Mr. Lesley about the calculus class his daughter was enrolled in. The teacher, he said, had a negative attitude toward all the students who weren't especially talented in mathematics. His daughter was an "A" student who worked hard and had always achieved well, but was finding calculus difficult. The parent accused her teacher, Mrs. Peerless, of not having enough patience with those students who needed more time for practice and review of the concepts. The parent was indignant and demanded that his daughter be switched to another

calculus class taught by Mrs. Lovejoy, who had a reputation for being enthusiastic about the subject and whose students always seemed to do well.

You are Mr. Lesley. How will you handle the situation?

Please think out loud as you consider this situation and your reactions to it.

Report Summary

The purpose of the report was to explore the meaning of grades assigned by different teachers of the same calculus course. The participants were 15 teachers, one from each high school in the board. All participants were highly qualified and experienced teachers of the course. The data collected were as follows:

1. a log giving the classroom activities each day and the time devoted to each;
2. a list of the homework and seatwork assigned each day;
3. a list of the criteria used to arrive at student grades for the course; and
4. grades assigned by the teachers to a common set of 20 final exams obtained from a class not involved in the study.

Great differences in practice were found. Among the teachers, the amount of homework and seatwork assigned varied greatly. When all activities listed in the logs were classified into one of six categories, variations in time spent in each activity were as follows: administration, from 0% to 5%; direct teaching, from 17% to 52%; student practice, from 8% to 47%; homework, from 11% to 43%; review, from 4% to 14%; and assessment, from 8% to 16%.

Teachers who placed a relatively high emphasis on practice activities tended to place a relatively low emphasis on review and assessment activities. Moreover, those with more class time available expended a smaller percentage of time on direct instruction and allocated a greater percentage to homework and practice.

From the reports of work assigned, the results showed that teachers varied widely in the number of questions assigned as homework—from under 500 to more than 1,600. Substantial differences were also found among teachers in content emphasis. For example, the number of questions on basic skills ranged from 40% to 75% of the total number of assigned questions; the number on proofs from 0% to 18%. Similar variations were found in the other content groups. When content of what was taught and what was tested were compared, there were irregular patterns in topic emphases. The emphasis on basic skills was less on the tests than the assignments, whereas the emphasis on other content groups was greater on the tests than on the assignments.

The analyses of the grading practices revealed large differences in the type of evidence collected by teachers for their grades. Examinations and term tests were the two main determinants of student grades. For the four classes (of 15) following an exemption policy, the majority of students took their only exam on material learned in the first half of the semester. For six other classes, the only exam was a final exam based on the entire semester's work. In the remaining classes, both a midterm and final exam were required. Across teachers, the final examination mark was weighted from 15%

to 40% of the student's final grade and, when a midcourse exam was administered, the resulting mark was weighted 9% to 30% of the final grade. Work in the first and second halves of the semester were weighted equally by five of the teachers, and the second half was weighted more heavily (70%) by the other 10.

The study of teacher marking of the common set of 20 exams revealed widely different expectations of student performance for marks awarded, as seen in substantially different marking schemes for the same questions and in the existence of "easy" and "hard" markers. Although there was substantial agreement among the teachers as to the relative importance of the examination questions and the relative quality of the student papers, there was substantial disagreement as to the absolute quality of the student papers.

To illustrate, three teachers assigned no paper a mark in the honors range, and one teacher assigned failing marks to seven papers; on the other hand, seven teachers assigned no paper a failing mark, and one teacher assigned marks of 80 or more to 10 of the 20 papers.

Appendix B *Value Statements*

Principal 1: Department—promote conformity; If not, they should be [common tests, exam], in other words one teacher sets the first test, another one sets the second one, and so on.

Principal 2: Department—promote conformity; It [the school] should [have an overall evaluation policy]. Department—study-like data; That isn't good enough [accept department's assessment]. Department—promote conformity; I don't want one class doing 8% and another class 40%.

Principal 3: Department—promote conformity; My particular philosophy is that parents and students have a right to feel that in one particular course where there are three or four or five different instructors that there is some continuity and some consistency of practice and evaluation. Department—promote conformity; While we are not trying to make robots of the teachers, we do feel that similar courses with similar evaluation practices should more or less resemble each other. Department—promote conformity; And while there will be some small difference in judgment, there is professional judgment, but that judgment should not be completely out of wack or based on entirely different tests that were given or different methods of evaluation.

Principal 4: School—transfer info; I think, first of all, you undermine confidence in your teachers, period, if you're willing to move them around like players on a chess board. Teacher—info; I think that's a principal's responsibility to see that the teachers' doing what they're supposed to be doing. Other; I'm not a believer in, first of all, undermining the teaching staff's strengths, period. Student—personal assistance; I am not a fan of personality conflicts. I don't allow it ... It's not something I encourage.

District—system policy; We feel different about final exams than they [other schools in board] do and I think you have to allow a certain amount of flexibility, but there has to be some check. District—avoid system exams; You

don't need examinations, maybe [in a school where focus is on business or tech] and you certainly don't need standardized testing in calculus for the number of kids you are dealing with. District—avoid system exams; And I think you ought to keep that independence in a school. District—avoid system exams; I think you ought to be able to be able to address the needs of your kids. Other; The grade 9 and 10 teachers should be dealing with the elementary feeder schools. Department—promote conformity; And I think this common set of exams, again it [discrepancy] shouldn't happen within a school. Department—promote conformity; There should be a marking scheme that's common. Department—promote conformity; There should be a common exam for the four sections in the school, if there are four teachers.

Principal 5: School—contain problem; And what you always want is to have the first avenue of appeal within the school.

Principal 6: Department—promote conformity; My feeling is that I would expect to see a relatively high degree of commonality in such things as courses of study, expectations of students, evaluation techniques, textbook being used, what is the purpose of evaluation, what methods are used, how much sharing of evaluation techniques, right through the whole year.

Principal 7: Other; I don't think you're going to get much support or much trust from teachers of any subject level working together to bring about betterment for the system if you're going to compare one school to another, or for that matter, even teachers within that school.

Principal 8: No value statements.

Principal 9: Department—don't promote; I don't think the solution is to demand that the daughter be switched to another calculus teacher. District—system policy; The department heads across the county should have some broad general standards ... Department—study-like data; and the department head in this particular school, when slicing that calculus pie into four pieces, should have some internal way to monitor the stresses in the courses—do we stress seatwork, do we stress the daily testing, do we have any assessment as opposed to evaluation, do we have make-up opportunities? ... master learning? ... chance to do [test] again? Department—promote conformity; There should be some understanding between these people so that we eliminate many of these differences ...

Principal 10: Other; I think you try and treat teachers ... treat them with respect.

Principal 11: No value statements.

Principal 12: Student; Whatever decision is made would have to be made in the best interest of the student, regardless of how the parent feels, and regardless of where the parent is coming from, and regardless really of what the teachers feel.

Principal 13: Teacher—attitude info; Sometimes that [teacher not being sensitive] happens knowingly and that probably bothers me more than anything else.

Principal 14: Department—promote conformity; I think in a common public educational system we need to be very very much aware of this and do whatever we possibly can to reduce that variance that could occur. Depart-

ment; You'll never, I don't think, get exact uniformity 'cause there are different teaching styles there.

Principal 15: Other; I believe that no matter where we are we can continue to improve and make things more effective. Department; I don't feel that everybody has to teach lock-step, that we're all doing all the same thing all of the time, because I think that certainly stymies individuality. Department—promote conformity; I also believe that there's got to be a core, that we are doing basically a similar thing, especially with respect to evaluation ... Promote conformity; I think it's only fair that there has to be a degree of continuity not only within a school but even within a system. Department—don't promote; I believe that we can't run away from a timetable from one person to the next. Student; I believe after thorough investigation if we have a personality conflict or a learning situation that is being stymied, then perhaps a change may be possible.

CLEMENT DASSA

Université de Montréal

From a Horizontal to a Vertical Method of Integrating Educational Diagnosis with Classroom Assessment

This paper is based on two Quebec studies which are part of an ongoing attempt to integrate educational diagnosis with class assessment. The underlying rationales, objectives, methods, and results of each research are described. Then, based on the limitations of the first study, as well as on new developments in the field of mastery learning, a mixed paradigm is outlined. It is argued that the unit of measurement should shift from the student to the teacher-student interaction, with the resultant proposal to change the methodology from a "horizontal" to a "vertical" one.

In the early 1980s, the Ministry of Education of the Province of Quebec began to implement a new policy concerning the evaluation of students in the primary, secondary, and college levels. Inspired by the recent developments in formative evaluation and mastery learning (Birzea, 1982; Block, 1971; Bloom, 1982; Bloom, Madaus, & Hastings, 1981), it emphasized the importance of student evaluation which became the primary responsibility of the teacher.

The ensuing reform of the principles and programs guiding the system made it necessary for the teacher to collect more information regarding students' learning and development and to do so more precisely than in the past. However, as this pedagogical reform was being gradually implemented, it became clear that the obvious problem of the quality of the information needed was compounded by the lack of integration of the "formative" tests with a global pedagogical strategy founded on the principles of formative evaluation.

This article addresses the problem of the integration of the theory and practice in this domain by presenting the objectives, methods, and results of two studies conducted in different Quebec school boards over a four year period beginning in 1983.

The common main goal of the two studies was to contribute to the elaboration of a model of diagnostic assessment destined to help guide the various educators (such as teachers, principals, and counselors) in the development of learning strategies based on formative evaluation. Each of the two studies focused on a specific research paradigm based on two different perspectives of classroom assessment, one centering on diagnostic testing as

Clement Dassa is an associate professor in the Faculty of Education. His research interests center on measurement and evaluation.

conceived by specialists, and the other focusing on classroom assessment as practiced by teachers.

The first study, *Elaboration et application d'un modèle de diagnostic pédagogique des difficultés d'apprentissage* intègre un processus d'évaluation formative,¹ focused essentially on assessing the efficiency of a heavily instrumented form of mastery learning approach. The second and current study, *Elaboration d'un modèle d'intégration du diagnostic pédagogique aux apprentissages scolaires*,² focuses on developing and validating a model of integration of educational diagnosis (diagnostic pédagogique) in the classroom setting.

The objectives of the first research were:

1. To analyze the different formative evaluation models;
2. To analyze the different methods of test development specific to educational diagnosis;
3. To develop a model of evaluation of the students learning difficulties;
4. To validate this model in a school setting; and
5. To develop a measurement strategy necessary to apply the model in a school setting.

The second study, which is presently in its initial stage, aims at:

1. Defining an effective research paradigm centered on the classroom;
2. Reassessing the previous model in the context of the *real* curriculum as practiced in the classroom;
3. Reformulating the model so that diagnostic assessment becomes more consistent with classroom constraints;
4. Focusing on long-term effects of the model; and
5. Exploring the feasibility of using classroom-based computerized diagnostic item banks.

Conceptual Framework

In both studies, formative evaluation is defined as any continuous evaluation practice intended to contribute to the improvement of learning in progress (Perrenoud, 1986). This definition has the advantage of not being limited to a given prescriptive model. In Canada, as in many other countries, we note in the past 20 years a growing number of official endorsements of the principle of formative evaluation. Whether in the form of published and widely diffused government regulations, as in Quebec for example (MEQ, 1979), or under different forms, the principle is widely recommended in policy statements.

However, moving from intent to practice has proved quite difficult. This problem has been addressed abundantly both from a theoretical and a practical point of view (Allal, 1983, 1988; Bloom, 1984, 1988; Cardinet, 1984, 1988; Scallon, 1988). At the core of all the mastery learning models one finds "feedback-corrective procedures" loops. The resultant "diagnosis" is often confounded with the various pedagogical interventions used; it is therefore intimately linked to both the fundamental cognitive processes involved in learning and to the interventions themselves. Hence the meaning and indeed

the definition of *diagnosis*, reflect the complexity of any pedagogical intervention.

In this article, we adopt an operational definition of educational diagnosis proposed in the context of computerized diagnostic assessment (Dassa, 1988) which allows one to take into account the applied nature of its integration to classroom assessment. In order to carry out an educational diagnosis one needs to:

1. Identify the error, that is, in the present classroom context, what the teacher identifies as an incorrect response;
2. Describe the nature of the error;
3. Describe the pedagogical context;
4. Propose a cause or a set of probable causes for the errors;
5. Attempt to link the cause(s) to a learning paradigm;
6. Propose corrective interventions coherent with both the teaching and technological approaches used and the classroom's ecology;
7. Assess the effectiveness of the corrective measures undertaken; and
8. Decide whether further remediation is needed: if not, end the diagnosis; if yes, decide on a further course of action.

The objectives of the first study were reached in three phases:

1. Determination of the state of the problem by means of extensive literature review and exploratory visits to numerous schools;
2. Elaboration of a diagnostic model and of a pedagogical guide for teachers; this phase also included a first empirical validation involving trial runs of all pedagogical material and tests;
3. Modification of the initial model according to the results obtained in the second phase; further validation of the model using a quasi-experimental design.

In this section we will focus mainly on the process of validation of the model (phase 3). Before presenting it in detail, we briefly summarize the model in its final form. As with most mastery learning models centering on corrective approaches, it is best described as an iterative process:

1. Determine an adequate structural hierarchy of a given learning unit;
2. Verify the affective and cognitive prerequisites;
3. Describe and analyze the difficulties students have and attempt to identify their causes;
4. Apply appropriate corrective measures;
5. Proceed with teaching according to a version of *pédagogie de maîtrise* (mastery learning according to Bonboir, 1970 and Allal, Cardinet, & Perrenoud, 1979);
6. Proceed with the next learning unit.

Essentially, the final validation consisted of comparing within a school setting the effects of various levels of integration of the diagnostic to teaching. The choice of these levels was set by the results of phase 1. In short, it was found that most schools use the Ministry's directives and documents in

an informal and uncontrolled way. Typically, the teacher is given lists of content-related objectives and then proceeds more or less as he or she wishes. In principle, teaching should take place according to the mastery learning approach as outlined in rather general terms in various Ministry documents available through the school board. Some teachers ignore these documents and teach in a traditional way. This situation characterized the “control” group. Two experimental conditions were then created to reflect: (a) the type of support the Ministry is considering offering (experimental 1); and (b) the maximum type of support that the Ministry, in conjunction with the local school board, would appear to be able to offer (experimental 2).

The teachers in the experimental groups were given teaching modules custom-made by the research team with the collaboration of the school boards to fit their programs and academic calendars. These learning units (LU) were developed according to the following criteria: (a) representativeness, the LU is a significant part of the program; (b) independence, the LU does not require highly specific prerequisites; (c) teachability, the LU should not present an unusual teaching problem; and (d) feasibility, the LU should be teachable within the normal school schedule.

The learning units consisted of teachers’ books complete with a guide detailing the objectives; the rationale for the sequence of subunits (lessons); detailed directives for students on how to use the material, namely, a series of “activities” used in each subunit; and a series of diagnostic tests. The experimental groups were differentiated by asking the teachers of the experimental 1 group to interpret the diagnostic tests by themselves, whereas the teachers of the experimental 2 group were given personal detailed instructions on the use of the diagnostic tests as well as all the assistance they required in interpreting and using their results.

The study was limited to grade 6; at this level Quebec schools tend to emphasize the consolidation of previously acquired topics. This contributed to increasing the independence and teachability of the LUs. Two LUs were especially developed for the experiment, one in French (as mother tongue) on homonyms and one in mathematics on the multiplication of numbers with decimal points.

Basically, the experiment was framed as a classical quasi-experimental design with repeated measures obtained on a large number of variables. The data originate from 33 of 36 grade 6 classes of the school board of Saint Jerome, a large town 50 km north of Montreal. The classes, comprising 854 students and 33 teachers, were assigned to 3 groups corresponding to the three levels of experimentation. The unit of assignment was the class and the system of random allocation controlled two factors: the location of the school (downtown or rural) and the socioeconomic status of the students attending the school (based on a socioeconomic study conducted by the school board).

All students were given a battery of pretests and posttests in mathematics and in French, a mathematics test on the prerequisite basic skills, a self-concept scale, and attitude scales for French and mathematics teaching. In addition, the students of the two experimental groups were given 10 diagnostic tests in mathematics and 6 in French; each lesson was followed by a short

paper-and-pencil quiz or by an oral activity (games, questions and answers, etc.). These tests covered the material of the 10 subunits in mathematics and the 17 (shorter) subunits in French. The experiment lasted about a month during which time the research team was present throughout the day in Saint Jerome mainly to respond to any request for assistance expressed by the teachers of the experimental 2 group and to keep track of relevant information pertaining to each class.

At the end of the experiment all participating teachers responded to a questionnaire. It focused on: (a) detailed reports on their use of pedagogical material and evaluation procedure used, (b) their reported degree of satisfaction with material and procedure, (c) their recommendation on material and evaluation procedure, and (d) their attitudes toward the different components of formative evaluation.

A univariate analysis of variance was performed separately on each variable of the pretests and posttests as well as on gain scores. This was followed by an analysis of covariance. The interactions between the affective and cognitive variables were investigated by a two-factor hierarchical design. Finally, the teachers' responses to the questionnaire were first analyzed in a descriptive way. These data were then combined with the information obtained by direct contact with the teachers during the experiment.

Summary of Results

It was found that the experimental groups generally presented higher achievement in French and in mathematics. Furthermore, the experimental 2 group was significantly different (higher) than the experimental 1 group in mathematics. In all cases there was a significant reduction of individual differences in the achievement variables for the experimental groups, as opposed to a nonsignificant (very slight) reduction for the control group. The effect size indices (Glass, McGaw, & Smith, 1981, p. 210) obtained for the experimental groups were equal to 0.56 for both French and mathematics.

The students' attitude had improved significantly only in the experimental 2 group and only in French. No significant correlation between attitude and academic achievement was found. In general, the LUs were well perceived by 70% of the teachers and students alike. However, about 30% of the teachers thought that the time required by the LUs was excessive. A majority (about 70%) of teachers thought that formative evaluation is useful and accepted the pedagogical consequences: verifying prerequisites, identifying difficulties, and remediating. However, it is worth noting that about 60% of the teachers thought that learning difficulties do not depend on the quality of teaching. Furthermore, about 25% of the respondents disagreed with the necessity for more personalized teaching.

Finally, the teachers of the experimental 1 group (who were not given personal detailed instructions on the use of the diagnostic tests) reported toward the end of the experimental period a change of attitude among some students with respect to the mode of testing. The gradual disenchantment of the students with the testing session, for both French and mathematics, seemed to express a rejection of the paper-and-pencil technology: Comments

of the type “Oh! not tests ... again!” seemed, according to some teachers, to be the typical way of greeting a session. Though the experimental 2 classes as a whole did not report such a “rejection” (to quote a teacher of experimental group 1), some of them, typically those who did not vary their instruction strategies, reported a gradual disenchantment of the students with the testing sessions.

Moving from Horizontal to Vertical Strategy

The limitations of Study 1 stem from a number of interrelated factors. The first is substantive in nature and is particular to all mastery learning models: Because the experiment lasted a little over four weeks, one should raise the question of the long-term effect of the application of the model studied. However, it should be noted that our study, which showed a 0.58 standard deviation effect, fared relatively well among the more comparable ones reported by Slavin (1987) in his meta-analysis of mastery learning research.

The second limiting factor concerns the method used. Study 1 was framed as a heavily instrumented traditional quasi-experiment based on the largest sample of subjects possible in order to ensure an adequate representativeness of the population of subjects. As a trade-off for generalizability with respect to the subjects, a severe limitation was put on the nature and scope of the curriculum as well as on the information pertaining to the teaching and evaluation naturally occurring in the classroom, the “real” curriculum.

The third and more general factor is epistemological in nature. Because the feedback-remediation loop is so central to the formative evaluation of learning in a teaching context, the errors made by the students have to be placed in their proper context. First, let us recall that an error is in practice defined as such essentially by the teacher. Hence errors express, at least partially, the type of relationship to knowledge that is gradually being built for the student. That is why corrective measures, in order to be effective, need to be related to the dynamic description of errors as well as to their fundamental underpinnings (statut de l'erreur). The corresponding diagnosis is therefore closely linked to the way the regulatory loop is put into practice or interpreted by the teachers, the school system, and for that matter by the researcher.

Essentially, the typical mastery learning approach starts with the formulation of objectives so specific that a large number of them are needed in order to satisfy curricular requirements. Cardinet (1988) has recently criticized the resultant “atomistic” evaluation approach. Even though the so-called “intermediate” (as opposed to “terminal”) objectives used in some school systems, such as in Quebec, may reduce the problem of inflation of objectives, the fact is that their choice remains essentially arbitrary. Neither the inflation nor the reduction of the objectives seems to be effective: When learning is measured precisely, then the number of objectives is typically very large and their choice is arbitrary; however, when a small number of terminal objectives are formulated, they tend to be too general to lend themselves to measurement.

A similar contradiction exists for setting the mastery criteria. Assessing an achievement as mastered or not mastered refers to the difficulty of correctly interpreting the results, of attributing a meaning to a given performance and thus of creating the proper frame of reference in terms of attempting to identify the cause of the difficulty experienced by the student. Whatever the reality of the evaluation performed, it is atomized in myriad measured entities, supposedly forming the basis of a diagnosis independent of the learning context and, a fortiori, from the classroom ecology.

Again, Cardinet (1988) showed how the prerequisites are, in this mechanistic interpretation of the regulatory loop, subject to the same criticism. This strikes at the core of the main operational device used in mastery learning: a sequence of objectives defining a hierarchy which is at best only suitable for certain types of learning. The nonlinearity implied by many language curricula as well as the need to restructure or construct certain mathematical representations are well known.

Furthermore, the diagnosis becomes problematic when conceived as an entity by itself, without any tie to the classroom. Indeed, its causal component could in some cases be useless. A simple example illustrates this point. In the same way as one could cure a bad case of flu by staying in bed and drinking plenty of fluids without ever identifying, or even attempting to identify, the nature or the physiological mechanism of the viral infection in order to cure it, a teacher could correct a difficulty by just repeating, and eventually changing, the didactic approach used. Of course, in some cases remediation without identifying the causes of the difficulties is next to impossible. Needless to say, persisting in this search raises the problem of allocating available resources in view of the gravity of the difficulty.

Finally, the problem of an unequivocal causal interpretation is not resolved by logically analyzing a given content of study, or even by doing so in reference to a specific learning theory. As noted by Cardinet (1988), a clear symptom (such as a difficulty with carrying numbers in arithmetic operations) may hide a subtler conceptual confusion (such as the function of numeration of position). When cut from its classroom context and from a broader subject matter theory, the simplistic sequence of "error, causal interpretation, and specific correction" seems to operate in a vacuum, because it excludes the interaction among the cognitive, affective, and social problems as they occur between the teacher and the student in the context of the classroom.

Method of Study 2: The Vertical Approach

From the previous section, it seems that corrective approaches stand a better chance of being effective, and the resultant learning of being closer to optimality, if regulation (as well as its ensuing measurement components) is conceived within the classroom context and in collaboration with the teachers. A new conception of regulation is thus needed, one that will more fully take into account the relevant affective variables pertaining to the teachers as well as to the students, one that will link the cognitive, affective, and social components of learning difficulties to the teachers' practices. With respect to

this conception of regulation, a new trend has appeared in Europe (compare, for example, Allal, 1988; Huberman, 1988). It proposes a less technological approach to evaluation, a more interactive and qualitative approach more in line with psychosocial analyses (compare, for example, Perrenoud, 1988). In North America, the need to tackle the evaluation of school achievements in a novel way, less dependent on testing and based on more thorough understanding of the classroom considered as a microworld, has led to recent developments in the way classroom assessment is studied (compare, for example, Stiggins & Bridgeford, 1985; Stiggins, 1988).

The second study discussed in this article may be situated at the junction of these two trends. The method used is based on a mixed paradigm (Van der Maren & Painchaud, 1984; Miles & Huberman, 1984). The nature and volume of the information required necessitated a compromise between the horizontal strategy retained for Study 1 (maximizing the number of students) and a more vertical approach aimed at maximizing the information from the basic unit of observation: the classroom.

The study is currently in its preliminary phase. Conceived as an exploratory research taking place in a high school of the Saint Jerome school board, its main specific objectives (related to the first objectives of the first study) can be summarized as follows: (a) to identify the relevant variables involved in managing the information necessary to classroom assessment; (b) to develop (or adapt) the instrumentation needed to assess the teachers perception, expectations, and practices related to classroom assessment; (c) to explore the feasibility and generalization of a model of on-site teacher training aimed at developing autonomy in performing diagnostic assessment.

A group of six Secondary II (second year of high school) mathematics teachers is currently participating in the study. After an initial period of negotiation and consultation with the school board, the head of the school, and the participating teachers, a model of intervention was elaborated and agreed on. Rather than proceeding from an initial *a priori* theoretical position, this method results from the confrontation of school constraints to the theoretical requirements of formative evaluation. Basically, the training centers on the preparation of an LU that will require about three weeks to teach (from mid-April to mid-May 1989).

The teachers are currently spending all their available time (about three hours every week), with the guidance of the researchers, elaborating pedagogical documents and tests, namely, interpreting the relevant terminal and intermediate objectives formulated by the Ministry of Education, formulating the domain of reference, devising a potentially useful (according to them) hierarchy in terms of the educational diagnosis, deciding on the didactic means to be used, elaborating their own assessment tools, clarifying the diagnostic interpretation that they plan to perform, and describing the remedial actions that they intend to implement.

The research team has already proceeded to a rather large collection of information relative to (a) the teachers' fundamental beliefs and attitudes regarding evaluation in general and classroom assessment in particular; (b)

the teachers' perceptions and expectations regarding evaluation practices in their own classroom, in their own school, and in the system at large; (c) the teachers' practices. The first two types of information were gathered by means of interviews and preliminary versions of beliefs and attitude scales; also, a comprehensive questionnaire is under development. Observation grids to be used in the classroom are being developed in an effort to better understand the way teachers really practice formative evaluation.

Concluding Remarks

The first study, conceived in the traditional positivist mode of investigation, yielded some interesting results concerning the efficiency of mastery learning. It showed that in the best possible case, making the most with pedagogical documents and tests prepared by outsiders, although custom made, had serious limitations. Even if it is unclear whether the sigma improvement would have persisted over a longer period of experimentation, the study did show, within the limitations discussed by Slavin (1987), that it pays to remediate difficulties. However, it also presented some intrinsic limitations which were discussed earlier. A new mixed approach has been developed to attempt to improve even further the efficiency of the regulation process, exemplified by the retroactive loop of mastery learning. Changing the focus of the study from the student as a statistical unit to the student-teacher relationship (one is tempted to say partnership) requires new research tools. The vertical approach, briefly described in this article, is presently being refined in the school. Its qualitative aspects, especially as seen by the participating teachers, bear some similarity to action research. By this, we mean that researchers and teachers alike are engaged in an ongoing, mutually agreed on exchange of information: The teachers are being trained and, in a way, the researchers too. However, the latter part of the exchange is only temporary and is due to the exploratory nature of the preliminary phase of the study. It is also genuinely confined by theoretical considerations relative to educational diagnosis and to its integration in a general framework of research. Once the present exploratory study ends, the second and third objectives described earlier should be pursued in a way that allows for generalization of the framework and its reformulation into a model.

Notes

1. Ajar, D., Dassa, C., and Scallon, G. Elaboration et application d'un modèle de diagnostic pédagogique des difficultés d'apprentissage intègre a un processus d'évaluation formative. FCAC-Quebec grant (1982-1985).
2. Dassa, C. Elaboration d'un modèle d'intégration du diagnostic pédagogique aux apprentissages scolaires. SSHRC-Canada grant 410-88-0886.

References

- Allal, L. (1983). Evaluation formative: Entre l'intuition et l'instrumentation. *Mesure et évaluation en éducation*, 5(6), 37-57.
- Allal, L. (1988). Vers un élargissement de la pédagogie de maîtrise: Processus de régulation interactive, rétroactive et proactive. In M. Huberman (Ed.), *Assurer la réussite des apprentissages scolaires?* (pp. 86-126). Neuchatel: Delachaux et Niestlé.
- Allal, L., Cardinet, J., Perrenoud, P. (Eds.). (1979). *L'évaluation formative dans un enseignement différencié*. Berne: Lang.

- Birzea, C. (1982). *La pédagogie du succès*. Paris: PUF.
- Block, J.H. (1971). *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston.
- Bloom, B.S. (1982). *Human characteristics and school learning*. New York: McGraw-Hill.
- Bloom, B.S. (1984). The 2 sigma problem, the search for methods of instruction as effective as one-to-one tutoring. *Educational Researcher*, 41(8), 4-17.
- Bloom, B.S. (1988). Le problème des deux sigmas: la recherche de méthodes d'enseignement en groupe aussi efficaces que le préceptorat. In M. Huberman (Ed.), *Assurer la réussite des apprentissages scolaires?* (pp. 45-85). Neuchatel: Delachaux et Niestlé.
- Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). *Evaluation to improve learning*. New York: McGraw-Hill.
- Bonboir, A. (1970). *La pédagogie corrective*. Paris: PUF.
- Cardinet, J. (1984). Le découpage des objectifs pédagogiques. *Mesure et évaluation en éducation*, 7(2), 5-48.
- Cardinet, J. (1988). La maîtrise communication réussite. In M. Huberman (Ed.), *Assurer la réussite des apprentissages scolaires?* (pp. 155-197). Neuchatel: Delachaux et Niestlé.
- Dassa, C. (1988). L'intégration du diagnostic pédagogique aux apprentissages scolaires: De la théorie la voie informatique. *Mesure et évaluation en éducation*, 11(1), 5-24.
- Glass, A.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Huberman, M. (1988). Les variantes de la pédagogie de maîtrise: Critiques et perspectives. In M. Huberman (Ed.), *Assurer la réussite des apprentissages scolaires?* (pp. 268-309). Neuchatel: Delachaux et Niestlé.
- MEQ. (1979). *Enoncé de politique et plan d'action*. Quebec: Ministère de l'éducation.
- Miles, B.M., & Huberman, A.M. (1984). *Qualitative data analysis*. Beverly Hills, CA: Sage.
- Perrenoud, P. (1986). *La part d'évaluation formative dans toute évaluation continue*. A paper presented at the congress Evaluer l'évaluation, Dijon.
- Perrenoud, P. (1988). La pédagogie de maîtrise: Une utopie rationaliste? In M. Huberman (Ed.), *Assurer la réussite des apprentissages scolaires?* (pp. 198-234). Neuchatel: Delachaux et Niestlé.
- Scallon, G. (1988). *L'évaluation formative* (Vols. 1-2). Quebec: Les Presses de l'Université Laval.
- Slavin, R.E. (1987). Mastery learning reconsidered. *Review of Educational Research*, 57(2), 175-213.
- Stiggins, R.J. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan*, 65, 363-368.
- Stiggins, R.J., & Bridgeford, N. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), 271-286.
- Van Der Maren, J.M., & Painchaud, G. (Eds.). (1984). *Actes du colloque sur les objets et méthodologies en recherche qualitative*. Montreal: FSE, Université de Montréal.

DAVID J. BATESON

The University of British Columbia

Measurement and Evaluation Practices of British Columbia Science Teachers¹

The 1986 British Columbia Science Assessment surveyed science teachers on many aspects of teaching practices, including measurement and evaluation practices. Science teachers at all grades depend most heavily on their own objective-type tests and activity write-ups for measurement information on which to base student evaluations. Oral measurement devices decline in use as students progress through the grades. Although the four goals of attitudes, skills and processes, knowledge, and critical thinking are equally important in the science curriculum, the emphasis in measurement and evaluation is on the knowledge goal, with attitudes being virtually ignored.

During the 1985-1986 school year the third assessment of science was undertaken in British Columbia. As part of that assessment, British Columbia science teachers were asked to respond to questionnaires about various facets of science teaching. Included in the questionnaires were items which dealt with measurement and evaluation practices. This article reports and discusses the results of those questions.

The subjects responding to the questionnaires were all practicing science teachers of grades 4, 7, and 10 in British Columbia schools. A 20% random sample of returned questionnaires at each grade level was drawn from the total populations for the purpose of detailed analysis. Table 1 reports the numbers of teachers in each grade level sample and their background data.

The ratio of male to female teachers at grade 4 is almost balanced, but as one progresses up the grades the ratio becomes unbalanced. By grade 10, over 90% of the teachers in the sample were male. In general, the teachers in the study were mature and well experienced in teaching, with over half the teachers having more than 10 years in the classroom.

All teachers in grades 4 and 7 were asked a question regarding the degree of emphasis they placed on certain types of measurement devices when deciding on a final evaluation for their students. The question was "How much emphasis do you place on each of the following in deriving a final evaluation for your students in science at this grade level?" Table 2 displays the teachers' responses to this question. The Weighted Use column of Table 2 was achieved by multiplying the percentages of teachers responding in the Little Use, Some Use, and Much Use categories by 1, 2, and 3 respectively, adding, and then dividing by 100.

Teachers were also provided the opportunity to suggest different types of measurement devices they may have used other than those suggested. None

David J. Bateson is an assistant professor in the Mathematics and Science Education Department, a research associate with the Educational Measurement Research Group, and an associate with the Centre for Applied Studies in Evaluation.

Table 1
Background Description of Teachers in the Study

Item	Category	Percentage of Teachers		
		Grade 4	Grade 7	Grade 10
<i>Number</i>		321	328	324
Gender	Male	51.1	75.7	90.1
	Female	48.9	24.3	9.9
Age	24 or less	1.0	1.6	0.6
	25-29	10.5	8.8	7.7
	30-34	17.5	17.9	15.8
	35-39	25.4	28.0	26.0
	40-49	33.3	36.8	39.0
	50 or more	12.4	6.9	10.8
Years of Teaching	1-2	4.4	3.7	4.6
	3-5	11.8	9.2	8.0
	6-10	20.2	22.5	20.4
	11-15	27.1	27.4	26.9
	over 15	36.4	37.2	40.1

was supplied. As can be seen from the table, teachers rely most heavily on their own objective-type tests, with activity/experiment write-ups and projects also receiving fairly heavy emphasis. Reports on science topics, anecdotal records, and subjective-type tests are the other devices which are used to a considerable extent. The only significant differences between grade 4 and 7 teachers were that grade 4 teachers use anecdotal records and oral tests to a higher degree than do grade 7 teachers. In addition, grade 7 teachers place more emphasis on activity/experiment write-ups. In general, however, grade 4 and 7 teachers are very much alike in the emphasis they place on particular measurement techniques for the purposes of gathering data to evaluate their students in science.

One surprising result was the lack of use of the Ministry of Education Classroom Achievement Tests. There are two tests at each grade level (Science Processes and Science Knowledge) available free in class sets to teachers in British Columbia. These tests are reliable, multiple-choice tests of 40 items each, with each item being referenced to the British Columbia elementary science curriculum. They are accompanied by teachers' manuals which include a section on standards set by practicing teachers in British Columbia. One can only speculate on why these tests do not receive more use given the teachers' reported heavy reliance on objective tests. Perhaps the teachers are not aware of their existence, in which case the Ministry should publicize them more. On the other hand, perhaps the confrontational attitude of "teachers versus the Ministry," which has existed in British Columbia for some time, has stimulated teachers to avoid anything to do with the Ministry, in which case perhaps distribution of the tests should be through

Table 2
Emphasis on Types of Measurement in Deriving Final Evaluation of Students
by Grade 4 and 7 Teachers (Percentage of Teachers)

<i>Type of Measurement</i>	<i>No (0) Emphasis</i>	<i>Little (1) Emphasis</i>	<i>Some (2) Emphasis</i>	<i>Much (3) Emphasis</i>	<i>Weighted Use (Rank)</i>
<i>Teacher-Made Objective Tests</i>					
Grade 4	1.0	2.9	48.1	48.1	2.43 (1)
Grade 7	0.0	2.2	43.3	54.4	2.52 (1)
<i>Projects</i>					
Grade 4	3.9	9.3	70.1	16.7	2.00 (2)
Grade 7	4.1	13.8	63.5	18.6	1.97 (3)
<i>Activity/Experiment Write-ups</i>					
Grade 4	5.2	16.4	62.3	16.1	1.89 (3)
Grade 7	4.1	11.1	57.6	27.2	2.08 (2)
<i>Reports on Topics in Science</i>					
Grade 4	6.1	14.5	68.1	11.3	1.85 (4)
Grade 7	6.0	16.6	63.6	13.8	1.85 (4)
<i>Anecdotal Records</i>					
Grade 4	7.3	23.6	58.3	10.8	1.72 (5)
Grade 7	15.8	31.8	46.2	6.2	1.43 (6)
<i>Subjective Tests (essay, paragraph, etc.)</i>					
Grade 4	13.8	21.7	57.6	6.9	1.58 (6)
Grade 7	9.3	22.8	57.6	10.3	1.69 (5)
<i>Oral Tests</i>					
Grade 4	35.0	27.6	32.7	4.8	1.07 (7)
Grade 7	47.4	31.1	19.9	1.7	0.76 (8)
<i>Individual Work Contracts</i>					
Grade 4	52.6	12.8	28.1	6.6	0.89 (8)
Grade 7	54.1	15.0	23.5	7.5	0.85 (7)
<i>Student Self-Reports</i>					
Grade 4	51.6	20.8	22.2	5.4	0.81 (9)
Grade 7	54.3	21.2	22.5	2.0	0.72 (9)
<i>Attendance</i>					
Grade 4	67.3	13.2	13.6	5.9	0.58 (10)
Grade 7	66.8	17.4	9.4	6.4	0.55 (11)
<i>Ministry of Education Classroom Achievement Tests</i>					
Grade 4	62.0	21.5	15.1	1.4	0.56 (11)
Grade 7	58.2	22.4	18.4	1.0	0.62 (10)

another agency such as the British Columbia Teachers' Federation lesson aids service.

Grade 10 teachers were asked a similar question and the results are reported in Table 3. Their question was, "How much importance do you place on each of the following in deriving a final evaluation for your students in Science 10?" The Weighted Use column in Table 3 was obtained by multiplying the proportion of teachers responding to the Important and Very Important categories by 1 and 2 respectively, adding, and dividing by 100.

Table 3
Importance of Types of Measurement in Deriving Final Evaluation of Students
by Grade 10 Teachers (Percentage of Teachers)

<i>Type of Measurement</i>	<i>Not (0) Important</i>	<i>(1) Important</i>	<i>Very (2) Important</i>	<i>Weighted Use (Rank)</i>
Teacher-Made Objective Tests	1.3	28.3	70.5	1.69 (1)
Laboratory Write-ups	3.1	59.4	37.4	1.34 (2)
Subjective Tests	13.6	56.2	30.2	1.17 (3)
Reports on Topics	19.7	63.5	16.8	0.97 (4)
Projects	21.7	61.2	17.1	0.95 (5)
Attendance	33.2	50.0	16.8	0.84 (6)
Anecdotal Records of Work Habits	29.0	63.0	7.9	0.79 (7)
Anecdotal Records of Achievement	41.4	39.1	19.5	0.78 (8)
Classroom Behavior	41.7	43.3	15.0	0.73 (9)
Standardized Objective Tests	43.0	42.7	14.3	0.71 (10)
Anecdotal Records of General Attitude in Class	47.7	48.3	4.0	0.56 (11)
Individual Work Contracts	72.3	21.3	6.4	0.34 (12)
Oral Tests	70.3	25.7	4.1	0.34 (13)
Student Self Reports	74.5	22.6	2.8	0.28 (14)

A similar pattern of importance to that seen from elementary teachers is apparent. There is a heavy reliance on teacher-made objective-type tests with laboratory write-ups also being important. Subjective tests appear to be more important at the secondary level than they are at elementary. These three devices were the only ones that received a mean importance of more than 1 (1=Important); however, reports and projects are also relatively highly considered. Attendance and classroom behavior become more important and oral tests become less important as the grade level increases. As was the case for elementary teachers, we see little reliance on standardized objective-type tests, although again there are several Ministry of Education Classroom Achievement Tests available free to teachers.

The *Junior Secondary Science Curriculum Guide and Resource Book* (British Columbia Ministry of Education, 1985) lists four goals for the junior secondary program: attitudes, skills and processes, knowledge, and higher level or critical thinking skills (creative, critical, and formal or abstract thinking). Further, it states that "all four goals should receive equal emphasis throughout the junior secondary science program" (p. 5). Grade 10 teachers were asked the question, "In the determination of the final letter grade for your students, what emphasis should be placed on each of the following?" Table 4 reports the results of this question.

As can be seen from Table 4, the goals are certainly not equally weighted when it comes to evaluation of students. It is possible that teachers are not aware that the goals are to be equally weighted. However, given the experience of teachers and the extent of inservice provided for implementation of

Table 4
Emphasis on the Various Goals for Final Evaluation by Grade 10 Teachers
(Percentage of Teachers)

Goal	Receives About							Mean Weight
	0%	10%	20%	30%	40%	50%	60%	
Attitudes	22.0	55.6	19.0	2.4	0.7	0.0	0.0	10.5%
Skills Processes	0.6	10.2	30.5	44.1	12.1	2.5	0.0	26.4%
Knowledge	0.0	0.6	7.9	23.1	27.8	21.5	19.0	41.9%
Critical Thinking	0.3	22.3	46.3	22.7	6.1	1.3	1.0	21.5%

the curriculum, this explanation for the lack of equal weight is improbable. It seems more probable that teachers are more familiar with, more comfortable with, and trust the techniques for measuring knowledge more than they do those for measuring other areas. Skills and processes can be measured through laboratory work, and Table 3 shows that teachers use laboratory write-ups to a considerable extent. On the other hand, the measurement techniques for assessing attitudes and critical thinking are, from my experience talking to teachers, neither familiar to teachers nor trusted by them. As such, it is probable that the lack of equivalence of the goals in final evaluations is at least partially due to a lack of trusted measurement information on these goals. Teachers need to be provided with good sample instruments and the skills and techniques necessary to develop, administer, and interpret measurement devices designed to assess areas other than the cognitive domain of learning.

Grade 10 teachers were asked whether they personally had the sole responsibility for the final evaluation of students or if some other measurements (school-wide or district-wide tests) played a part in determining a student's final letter grade. About two thirds of the teachers indicated that they had either or both a school-wide and/or district-wide test, while about one third said they had the sole responsibility. They were then asked what their preference would be as far as external examinations were concerned and what weight should be given to an external exam. Table 5 reports the results of this question.

In general, teachers are much more favorable toward a school-wide exam than to a district-wide or provincial examination, with only between one quarter and one fifth of the teachers agreeing with the latter two. Given that teachers put little emphasis on standardized objective examinations as displayed in Table 3, this is an expected finding. If external examinations were to come into play, the most popular weight would be about 20% which, from a measurement perspective, is not unreasonable.

The teachers were also asked two questions regarding the distribution of marks they assigned to students. The first question is whether they used a preset distribution of marks for their science students. Over one third of the teachers responded yes, they did use a preset distribution of marks! This

Table 5
Grade 10 Teachers' Preferences for External Examinations
(Percentage of Teachers)

Type of Measure	No	Yes, worth:				
		10%	20%	30%	40%	50%
School-Wide Exam	20.6	8.2	50.4	12.8	3.9	4.3
District-Wide Exam	74.5	9.6	11.6	2.4	0.4	1.6
Provincial Exam	77.1	5.5	8.5	4.8	1.1	3.0

finding is alarming given the extreme differences in ability and achievement that occur from class to class, from school to school, and from year to year.

The second question asked teachers to provide their expectations of mark distributions for their science classes. Thirteen percent of teachers said that they had no expectations for any given class. The responses from the remaining 87% are illustrated in Table 6.

The Mean column was obtained by multiplying the percentage of teachers by the midpoint of the range, adding, and dividing by the total percentage of teachers responding. The resulting expected distribution of letter grades appears to be very traditional and conforms quite closely to the reported results over the last few years for English 12, the only government examinable subject in Grade 12 that all students must take.

The results of this study contain few surprises, although the lack of use of available, curriculum-referenced, quality objective tests is a finding which should cause some concern. Teachers in British Columbia are probably not different from teachers in other provinces. They receive little training in evaluation in their preservice education, often only part of one course, and

Table 6
Grade 10 Teachers' Expectations of Student Letter-Grades
(Percentage of Teachers)

Letter Grade	Percentage of Students					Mean
	0-5%	6-15%	16-30%	31-50%	51-100%	
A (Excellent)	56.5	37.7	4.5	1.0	0.3	7.9%
B (Very Good)	5.8	59.7	31.4	3.1	0.0	10.5%
C+ (Satisfactory)	3.2	28.4	60.3	7.4	0.7	23.1%
C (Satisfactory)	1.1	15.2	55.3	26.2	2.1	29.8%
P (Pass)	11.4	44.8	31.0	10.0	2.8	20.6%
F (Fail)	57.4	35.2	6.0	1.1	0.4	8.2%
No Expectations			13%			

few teachers elect measurement and/or evaluation as a component of their ongoing inservice education. As McLean (1985) points out, evaluation is more a craft than a profession; teachers measure and evaluate more through "folk knowledge" than from a sound theoretical and practical base (pp. 32-34). At present, teachers appear to depend heavily on evaluation in the cognitive domain, an area rich in folk knowledge. Other areas are severely underutilized. The theoretical knowledge and practical techniques for measuring all domains of education exist. However, some method must be developed to provide teachers with good examples of measurement instruments and with the existing knowledge and the practical skills necessary to use alternate, appropriate measurement and evaluation techniques for the assessment of particularly the affective and psychomotor domains. The award-winning publication *Evaluation Techniques: A Handbook for Teachers* (British Columbia Primary Teachers' Association, 1985) has been well received and has had considerable impact on primary teachers in British Columbia. It might serve as one model to be considered. Perhaps an extended document containing concepts such as those expressed by Cornett (1982) could be produced for all grade levels. In any case, the challenge to the Canadian measurement community is to have the knowledge and techniques that exist put into practice.

Note

1. The data used in this study were collected under contract to the British Columbia Ministry of Education and were first reported in the British Columbia Science Assessment Reports (Bateson, Anderson, Dale, McConnell, & Rutherford, 1986a; Bateson, Anderson, Dale, McConnell, & Rutherford, 1986b). The author is grateful for their permission and assistance.

References

- Bateson, D.J., Anderson, J.O., Dale, T., McConnell, V., & Rutherford, C. (1986a). *The 1986 British Columbia science assessment: General report*. Victoria: British Columbia Ministry of Education.
- Bateson, D.J., Anderson, J.O., Dale, T., McConnell, V., & Rutherford, C. (1986b). *The 1986 British Columbia science assessment: Summary report*. Victoria: British Columbia Ministry of Education.
- British Columbia Ministry of Education. (1985). *Junior secondary science curriculum guide and resource book*. Victoria: Province of British Columbia.
- British Columbia Primary Teachers' Association. (1985). *Evaluation techniques: A handbook for teachers*. Vancouver: British Columbia Primary Teachers Association.
- Cornett, J.D. (1982, September) Alternatives to paper-and-pencil testing. *NASSP Bulletin*, 66, 44-46.
- McLean, L.D. (1985). *The craft of student evaluation in Canada*. Toronto: Canadian Education Association.

W. TODD ROGERS

University of Alberta

Current Educational Climate in Relation to Testing

As in other jurisdictions, there is strong support for testing and assessment in British Columbia. Taken as a whole, the data and information obtained in the Let's Talk About Schools study and the Royal Commission on Education, completed respectively in 1985 and 1988, suggest that there is far more support for new and expanded examination, assessment, and evaluation programs. Despite this apparent popularity, members of the measurement and evaluation community must be mindful of the concerns raised by those who question the "overuse" of tests and continually monitor the way tests are developed and used.

Communiqué

Regina, September 27, 1988

FOR IMMEDIATE RELEASE

EDUCATION MINISTERS DISCUSS NATIONAL ISSUES

Provincial and territorial ministers responsible for education concluded a two-day meeting of the Council of Ministers of Education, Canada (CMEC), in Regina, Saskatchewan, today.

The ministers agreed that provinces will coordinate the assessment of the quality and effectiveness of their education systems. A proposal for a three-phase implementation plan was approved. The assessment will examine participation and graduation rates and, in the final phase, a means will be developed for assessing achievement levels of 13 year olds across the country in literacy and numeracy.

PROVINCE OF BRITISH COLUMBIA NEWS RELEASE

MINISTRY OF EDUCATION

FOR IMMEDIATE RELEASE

NEWS RELEASE #06-89

DATE: JANUARY 27, 1989

MAJOR CHANGES TO EDUCATION ANNOUNCED

Victoria, B.C.—Minister of Education Tony Brummet today announced major new policy directions for education in British Columbia.

Brummet reported that government has accepted the vast majority of the recommendations of a recent Royal Commission of Education headed by Barry Sullivan....

Sullivan's report pointed out that if our education system is to maintain its relevance and prepare our children for the future, it must change to meet these new challenges.

Major changes announced by Brummet include a new provincial curriculum structure that will be phased into schools over the next ten years....

Expanded assessment and accountability systems aimed at monitoring the progress of individual students and reporting to the public will also be developed. The Grade 12 provincial examination system will be expanded....

Todd Rogers is a professor of educational psychology. His research interests are measurement, evaluation, and research methods.

The above quotations reflect the growing concern with the quality of schools and the future of those who attend schools, the students. Briefly, parents concerned with the progress of their own children and educators responsible for the provision and delivery of quality educational programs express uneasiness about the ability of today's schools to adequately meet the needs of students and to prepare students for what can best be described as a rapidly changing technological future. Governments, in response to these concerns, are directing increasing proportions of their budgets toward education and its improvement.

That calls for new and expanded testing have been made is predictable and reflects what might best be called a theme for greater accountability. Such a theme is made evident in public and professional support for standards and examinations and in a desire to monitor student performance. Clearly there is a belief among the public and professionals taken as a whole that standards and testing are useful ways to monitor and evaluate school effectiveness and, thereby, to ensure that accountability with respect to quality exists within the system.¹ Politicians responsible for allocating funds for education want to know if they are getting their money's worth and turn to testing as the way to find out.

Every large education reform of the past few years has called for either new tests or expanded use of existing ones. In Britain, for example, new national tests are a key tool in the government's effort to enact far-reaching changes in that nation's education system. Central to this reform is a unified national curriculum backed by a series of tests administered to children at the ages of 7, 11, 14, and 16. In its recent deliberations about the bill to reauthorize federal elementary and secondary education programs, the United States Senate added an amendment that calls for a substantial expansion and reshaping of the federal program that measures educational progress, the National Assessment of Educational Progress. The stated purpose of the Senate amendment was to improve the effectiveness of schools by making objective information about student performance in a variety of learning areas (reading, writing, mathematics, science, history, geography, and civics) available to policymakers at various levels.

This purpose closely parallels the statements of purpose that can be found in the rash of legislation mandating new or expanded testing enacted in various states over the last five or so years in response to the somewhat scathing indictment of United States schools by the National Commission on Excellence in Education (1983). Petrie (1987) summarized this situation by indicating that it is not "too much of an exaggeration to say that evaluation and testing have become the engine for implementing educational policy" (p. 175).

The call by the British Columbia Minister of Education (Brummet, 1989) for increased assessment and testing as part of the new curriculum structure to be implemented over the next 10 years is yet another example of this trend. The pertinent policy statements and their rationales are:

1. The regular Grade 12 provincial examination system will be expanded, and student performance on Grade 12 scholarship examinations will be given greater recognition.

It is important that students have the opportunity to demonstrate their competence, in comparison to provincial standards, for as many areas as possible of the provincial curriculum which they have studied in their final year of school. Outstanding scholarship must also be recognized more consistently.

Where they do not now exist, regular provincial examinations will be developed for Grade 12 courses in the Humanities area of the current provincial curriculum. As the new provincial curriculum is developed, all Grade 12 courses in the Humanities and Sciences areas will have provincial examinations.

Students will be asked to demonstrate their competence in the other areas of the new provincial curriculum, Practical and Fine Arts, prior to graduation.

Effective in the 1989-90 school year, provincial examinations will be given a 40% weighting in determining a student's final course mark in examinable Grade 12 courses. This is a reduction from the present system in which provincial examinations are given a 50% weighting. The change reflects a greater emphasis on teachers' assessment of the classroom performance of students in their final year of school.

2. There will be increased emphasis on regular assessment of student performance and evaluation of program effectiveness.

Vitally important to the continual improvement of student programs is careful monitoring of how well students are learning in comparison to the intended outcomes of instruction. Current efforts of the system in this regard need to be expanded.

A variety of new assessment and program evaluation procedures will be developed for use at the classroom, school, and provincial levels. These procedures will facilitate comparing program results to target levels of performance and provincial standards.

The current Provincial Learning Assessment Program will be expanded and adjusted to focus on the objectives of the new provincial curriculum.

As an additional measure of program effectiveness, the province will continue to take advantage of opportunities to participate with other provinces and countries in comparison studies of student performance.

To provide relevant information efficiently for purposes of monitoring outcomes at all levels of the school system, a new student data collection and processing system will be implemented in cooperation with districts and schools.

3. Accountability and reporting procedures at all levels of the system will be improved.

Students need to be informed about their strengths as well as about areas that require further work. Parents, the local community and the general public expect, and have a right to receive, information about the operation of the system on a regular basis and in an understandable form. The strengths of the current reporting procedures need to be developed further.

Students will be provided with continuous feedback from teachers about the skills they have mastered and about those that require further development.

Parents will receive frequent and clearly understandable reports from schools on student progress and competence.

The community will be provided with annual reports from the school board on the performance of schools in the district, in relation to provincial and district objectives and standards. The reports will reference key indicators, based on information provided by the Ministry.

The Legislative Assembly and the people of the province will receive annual reports from the Ministry showing the progress made toward achieving the goals of education and the desired attributes, as outlined in the mandate for the system. (pp. 14-16)

Further evidence of the increased role assessment and evaluation are to play in the new provincial curriculum can be seen in the ungraded program for the first four years of schooling. During the first of these years (formerly kindergarten),

the individual child's developmental needs can be assessed and appropriate program adjustments made.... Schools will determine the program for each child based upon the child's needs, local priorities, and the capabilities of the school. Children can move forward at their own pace. Some students may complete the program in less than four years, and some may take longer. (Brummet, 1989, p. 12)

To underscore the importance of these assessment and evaluation initiatives, the Ministry set September 1989 as the deadline for completing the planning necessary for the implementation of new and expanded assessment and examination procedures.

Analysis of the Ministry's assessment and evaluation policy suggests that evaluation and testing will play instrumental roles in the new provincial program to be implemented in British Columbia. This policy reflects what Petrie (1987) has called the engine of educational policy and is in agreement with testing policies in other jurisdictions like Great Britain where major educational reform is taking place. But is there support for such a role in British Columbia?

Generally, the answer is yes. Support comes from two major educational studies completed in British Columbia during the last four years: *Let's Talk About Schools* (Ministry of Education, 1985) and the Royal Commission on Education (Sullivan, 1988).

Let's Talk About Schools

The intent of *Let's Talk About Schools* (Ministry of Education, 1985) was to provide a valid and comprehensive portrayal of British Columbians' opinions about and preferences for their schools. Unlike a Royal Commission, no recommendations were to be formulated.

This study began in July 1984 when the then Minister of Education assembled an Advisory Committee to prepare a background paper which would help stimulate public and professional discussion of some of the major issues and questions facing schools now and in the years ahead. Of the 41 discussion questions formulated by the Committee, three organized under the heading Standards of Achievement in Schools were related to assessment and evaluation:

1. What standards should be established for all schools, public and otherwise? Should standards be established for all grades and all subjects? Who should establish these standards? Should standards be monitored? How? By whom?
2. To what extent should province-wide examinations exist? What purposes can such examinations serve?
3. To what extent should entrance examinations set by postsecondary institutions determine secondary school curricula and standards? (Ministry of Education, 1985, Booklet 5, p. 12)

Data Base

February 1 to mid-March 1985 was the period set aside for public and professional discussion of the 41 questions plus any other issues that participants wished to raise. Altogether 3,525 individual responses from parents and members of the public, 969 individual responses from education profes-

sionals (teachers, administrators, school trustees, teacher educators), 482 letters, 316 group submissions, and 10 petitions (signed by a total of 5,110 people) were received. During this same period, 567 meetings with an estimated total attendance of 15,575 were held throughout the province.

Analysis

The coding system developed to analyze responses was generated from the answers provided by the respondents to the discussion questions. For each of the 41 questions, the statements made were grouped according to common elements. Themes were then extracted and designated as “coding components.” Codes reflecting the variation within each component were then formulated. For example, analysis of the answers given in response to Question 2: “To what extent should province-wide examinations exist? What purposes can such examinations serve?” led to the identification of seven components. The first addressed the issue “Should there be province-wide examinations?” The corresponding codes were 1=yes, 2=no, and 3=partly. The second component, with its 14 codes, captured the functions and purposes held by the respondents for these tests. The third component with its four categories was necessary to reflect the negative consequences of testing identified by respondents opposed to province-wide examinations. Components 4 and 5 picked up those respondents who identified the grade(s) and/or subject(s) for province-wide testing. Some respondents proposed qualifications regarding the types of students to be tested, thus the need for a sixth component. Finally, it was necessary to introduce a seventh component to capture statements made about the degree to which province-wide examinations should count toward final or year-end grades.

Table 1 presents a summary of the coded responses to the three achievement standards questions. The first column lists the components with their codes. The next two columns contain the percentages of parents and general public (column 2) and education professionals (column 3) for each of the codes. The corresponding percentages for meetings, letter writers, and groups appear in columns 4 to 6. In all cases the percentages are based on the number of respondents, *N*, coded in each component. Percentage values are not reported for the code “Other” which includes responses made by fewer than 5% of the respondents for that component.

Findings

As shown in Table 1, for the most part the component percentage distributions across the five reporting groups are comparable. In order to simplify reporting in the presentation of highlights which follow, the term respondent is used rather than separately listing the five samples in these cases. Where differences occur, the samples are identified by name.

1. What standards ... ?

More than 90% of the respondents agreed that achievement standards should be established for all forms of schooling. Approximately half favored provincial standards, while national standards were preferred by one third of the general public; by one fifth of the education professionals, letter writers,

Table 1
Standards of Achievement in Schools

	GENERAL PUBLIC RESPONSE DOCUMENTS	EDUCATIONAL PROFESSIONALS RESPONSE DOCUMENTS	MEETINGS	LETTER WRITERS	GROUPS
1	What standards should be established for all schools, public and otherwise? Should standards be established for all grades and all subjects? Who should establish these standards? Should standards be monitored? How? By whom?				
1.1	N = 2514	N = 676	N = 344	N = 90	N = 91
Yes	95.0	92.5	98.0	94.4	96.7
No	2.0	2.7	0.6	1.1	2.2
Other (public only, private only, not possible to set standards)					
1.2	N = 1022	N = 247	N = 178	N = 42	N = 50
What should be the reference group for standards?					
National	34.8	18.6	33.7	23.8	18.0
Provincial	53.6	50.6	57.9	50.0	48.0
District	3.5	13.4	6.7	2.4	6.0
School	4.9	7.3	2.2	0.0	10.0
Individual Student	12.7	22.7	8.4	19.0	32.0
Other Suggestions (international, classroom)					
1.3	N = 1106	N = 218	N = 102	N = 24	N = 36
Which grades should standards apply to?					
All Grades	82.3	67.4	65.7	62.5	75.0
Elementary	6.5	8.7	9.8	12.5	8.3
Junior Secondary	6.0	7.3	8.8	12.5	5.6
Senior Secondary	9.1	16.5	11.8	12.5	5.6
1.4	N = 1074	N = 212	N = 95	N = 23	N = 34
Which subjects should standards apply to?					
All Subjects	59.0	44.3	31.6	30.4	58.8
Core Subjects	38.5	52.4	62.1	60.9	41.2
Other Subjects (special education subjects, vocational subjects)					
1.5	N = 935	N = 254	N = 133	N = 56	N = 50
What is the basis for the standards?					
Average	6.1	5.5	1.5	1.8	0.0
Higher Than Average	8.4	7.1	6.8	7.1	2.0
Mastery (when specifically stated)	14.7	14.2	12.8	8.9	4.0
Minimum Competence (functionally literate)	22.9	24.8	24.1	16.1	46.0
Honours/Pass/Fail	4.9	5.5	6.0	5.4	6.0
Excellence (highest possible)	33.5	31.1	45.1	66.1	40.0
Other (system based on percentages, guide lines only, present system is fine as is)					
1.6	N = 1548	N = 399	N = 118	N = 31	N = 57
Who should be involved in setting standards?					
Set. Consu.*		Set. Consu.	Set. Consu.	Set. Consu.	Set. Consu.
Province (Ministry of Education)	53.0 1.8	44.6 4.0	44.1 1.7	61.3 0.0	50.9 3.5
School Board (district, central administration)	13.9 3.5	18.8 4.5	9.3 7.6	3.2 0.0	19.3 10.5
Professionals (schools, BCTF, teachers, principals, experts, school administrators)	33.6 7.0	38.3 11.0	32.2 11.9	19.4 6.5	26.3 19.3
Parents	10.8 4.8	9.0 3.5	6.8 5.9	6.5 3.2	17.2 19.3
Broad-based Committee (committee with mixed range of representatives)	5.6 0.4	7.3 0.3	8.5 0.8	3.2 0.0	5.3 0.0
Other Suggestions (Provincial specialist associations, local specialist associations, private schools set own, pupils)					
*NOTE: Set. — Set the standards. Consu. — Consultant to assist in setting standards					
1.7	N = 1311	N = 301	N = 93	N = 39	N = 56
Should standards be monitored?					
Yes	97.1	96.0	97.8	89.7	96.4
Other (no, not really, teachers should be monitored)					
1.8	N = 763	N = 162	N = 51	N = 19	N = 34
How should they be monitored?					
Government Exams	65.4	58.6	58.8	68.4	47.1
Teacher Tests (exams, tests, unspecified)	28.7	30.9	45.1	26.3	41.2
Classroom Observation (class visits, general assessment)	10.2	14.8	7.8	5.3	23.5
Learning Assessments (Canadian Achievement Tests, Provincial Learning Assessment Program)	8.3	9.9	3.9	15.8	17.6
1.9	N = 970	N = 235	N = 50	N = 21	N = 40
By whom should they be monitored?					
Provincial Officials (Ministry)	44.5	39.6	40.0	61.9	60.0
District Officials (Superintendent, district administrators)	13.1	22.1	22.0	4.8	12.5
School Board	20.3	17.0	10.0	23.8	17.5
Principals (school administrators)	11.0	11.5	20.0	9.5	10.0
Teachers (BCTF)	22.7	28.9	26.0	33.3	30.0
Independent Committee (broad based committee)	7.2	5.5	8.0	0.0	7.5
Parents	13.0	11.1	4.0	9.5	5.0

Table 1 (continued)

2	To what extent should Province-wide examinations exist? What purposes can such examinations serve?					
2.1	Should there be Province-wide examinations?	N = 2459	N = 675	N = 280	N = 68	N = 76
	Yes	76.5	68.3	89.6	73.5	77.6
	No	14.9	19.0	5.0	19.1	14.5
	Qualified yes ("Yes, but ...")	7.0	11.3	4.3	7.4	7.9
	Other (uncertain, not sure)					
2.2	The following purposes were suggested by the respondents who favoured (yes and qualified yes) Province-wide testing:	N = 1647	N = 379	N = 150	N = 41	N = 52
	Use with Standards (way to see if standards and guidelines are being met)	45.7	43.3	42.0	61.0	38.5
	Normative Comparisons (how we stand relative to others)	5.8	8.2	10.0	4.9	3.8
	Ensure Uniformity (guarantee equality of education throughout the Province/districts, same education for all in B.C.)	20.0	12.4	22.7	24.4	23.1
	Ensure the Curriculum is Being Followed	11.7	12.1	12.0	12.2	17.3
	Measure Achievement	12.9	14.8	14.0	19.5	17.3
	Provide Feedback at Student Level	8.5	13.5	13.3	7.3	23.1
	Assist with Student Placement	10.9	11.4	5.3	9.8	17.3
	Assist Evaluate/Monitor Students	20.6	16.6	10.6	19.5	25.0
	Assist Evaluate/Monitor Teachers	20.1	14.5	17.3	31.7	15.4
	Assist Evaluate/Monitor Schools	8.6	5.5	3.3	4.9	1.9
	Other Purposes (provide feedback to parents, provide feedback to employers, evaluate/monitor principals and district staff)					
2.3	The following reasons were given by the respondents who opposed Province-wide examinations:	N = 186	N = 74			
	Inhibit Learning (an ineffective way of testing, too much stress on students, negative effect on testing practices)	48.4	55.4	Insufficient number of respondents to conduct analysis.	Insufficient number of respondents to conduct analysis.	Insufficient number of respondents to conduct analysis.
	Teachers Teach to Test	26.9	37.8			
	Expensive/Waste of Time	32.3	25.7			
	District/School Responsibility	10.8	12.2			
2.4	Which grade should be tested?	N = 838	N = 211	N = 102	N = 14	N = 36
	Grade 1	23.2	12.3	0.0	0.0	0.0
	Grade 2	23.6	12.3	21.6	0.0	0.0
	Grade 3	27.2	18.0	24.5	7.1	30.5
	Grade 4	27.3	19.0	32.4	0.0	25.0
	Grade 5	25.7	14.7	25.5	0.0	25.0
	Grade 6	31.6	20.4	28.4	7.1	27.7
	Grade 7	41.4	30.8	46.1	21.4	36.1
	Grade 8	42.4	28.0	46.1	14.3	36.1
	Grade 9	48.9	27.0	51.0	21.4	38.8
	Grade 10	56.0	41.7	64.7	21.4	41.6
	Grade 11	57.5	40.8	56.9	35.7	50.0
	Grade 12	90.7	91.0	90.2	71.4	88.8
2.5	Which subjects should be examined?	N = 352	N = 93	N = 34		N = 12
	All Subjects	22.2	14.0	29.4	Insufficient number of respondents to conduct analysis	0.0
	Academic and Core Subjects	69.3	81.8	67.6		100.0
	Other (special education, vocational)					
2.6	How should the examinations count toward a final grade? How should they be weighted?	N = 401	N = 156	N = 107		N = 20
	Less Than 50%	40.1	55.1	49.5	Insufficient number of respondents to conduct analysis.	50.0
	50%	9.7	9.0	12.2		0.0
	More Than 50%	6.0	3.2	5.6		0.0
	Partial Only	30.4	19.9	4.6		25.0
	Not Final (do not count as part of final mark at all)	10.0	7.1	28.0		25.0
3	To what extent should entrance requirements set by post-secondary institutions determine secondary school curricula and standards?					
3.1	Should post-secondary entrance requirements determine secondary curricula and standards?	N = 1714	N = 447	N = 91	N = 30	N = 42
	Yes	78.9	68.9	70.4	83.3	71.4
	No	18.3	28.6	26.4	16.7	28.6
	Other (uncertain, not sure)					

and groups; and by one third of the meetings. Approximately 20% of the education professionals and 30% of the groups indicated that standards should be set at the individual level (Ministry of Education, 1985, 1.1, 1.2).

A majority of respondents indicated that standards should be established for all grades. They were less certain about which subjects should have standards. Six out of 10 meetings, letter writers, and groups favored restricting standards to core subjects, while six out of 10 public respondents suggested standards should be applied to all subjects. Education professionals were more evenly split (1.3, 1.4).

Generally, respondents preferred maximum performance over minimum competency (1.5).

Preferences of the respondents concerning who should establish standards were, in order, provincial officials, education professionals, and school board trustees (1.6).

Respondents wanted the standards monitored and suggested that this responsibility be given, in order of preference, to provincial officials, teachers and district officials, and school board members. Provincial examinations and classroom tests were preferred methods for such monitoring (1.7, 1.8, 1.9).

2. To what extent should province-wide ... ?

Approximately 80% of the general public, education professionals, letter writers and groups, as well as 90% of the meetings, called for province-wide examinations. Of these respondents, more than 90% advocated Provincial testing at Grade 12. Further, beginning with Grade 1, the percentages tended to increase from one grade to the next, with marked increases at the grade levels corresponding to the end of elementary and junior secondary schools. Overall, the public placed greater emphasis on testing than did education professionals (2.1, 2.4).

The most frequently cited reasons given for provincial exams by their supporters were to uphold standards; to ensure uniformity among districts; to measure, evaluate, and monitor student achievement; and to evaluate and monitor teacher performance. Those opposed to province-wide tests stated that such tests "inhibit learning," "lead teachers to teach to the tests," and "are expensive and a waste of time" (2.2, 2.3).

Slightly more than two thirds of the general public and meetings, 82% of the professionals, and all groups suggested provincial tests should be administered in academic and core subjects. Of those who commented on the degree to which province-wide tests should count toward final grades, the majority indicated less than 50% (2.5, 2.6).

3. To what extent should entrance ... ?

Approximately 80% of the public and letter writers felt postsecondary entrance requirements should determine secondary school curricula and standards. About 70% of the education professionals and groups also endorsed this view (3.1).

Royal Commission on Education

The Royal Commission study of British Columbia schools (Sullivan, 1988) was initiated in March 1987 and completed in July 1988. Its tasks were to:

1. Report on the state of British Columbia schools in both the public and nonpublic sectors;
2. Identify areas of difficulty or concern as part of an assessment of the present and future needs of the system;
3. Identify the kinds of preparation programs school graduates will require to meet the challenges of the next two decades;
4. Determine what school programs and administrative processes should be implemented to deal with current and emerging problems; and
5. Recommend some responsible paths of action based on comprehensive understandings of the system's characteristics, priorities, and the social and economic contexts in which schools operate (p. 4).

Data Base

During the period set aside for data and information gathering (May 1987 through March 1988), the Commissioner and his staff visited 139 schools in 89 communities, participated in 23 student assemblies, and held 66 public hearings. In addition to these public meetings, 54 meetings were held with teachers alone. In all, 2,099 written and 231 oral submissions were received. Over the course of the data collection period, additional meetings were held with representatives of government agencies and the major provincial educational organizations to clarify points made in written submissions received beforehand. In addition to the information presented in the oral and written submissions, the Commission sought and obtained information from a variety of educational reports from other provinces and from jurisdictions outside of Canada, as well as from journal articles, government reports and commissioned position papers, and research conducted by Commission staff.

In light of concerns with parents' rights and responsibilities, the preschool experience of children, and the articulation between school and work, three working groups were appointed to advise the Commission on these issues. The Commission was similarly concerned that the views of students be heard, especially those just about to complete school. Accordingly, a student task force consisting of 54 senior secondary students from throughout the province was established for this purpose.

During the course of the Commission, the Commissioner was assisted by six research teams within six broad areas related to education in the province.² Each team systematically analyzed all the information described above. As well, the team members attended public hearings, visited schools, participated in conferences, reviewed research articles and reports pertinent to their own areas, consulted experts, commissioned their own position papers, and conducted their own research. At the completion of its work, each team prepared a technical report containing the results of its analysis and a series of suggestions designed to close the gap between "what is" and "what ought to be."

Following submission of the six team reports, representatives from each team systematically analyzed all reports to develop, in concert, an integrated set of suggestions for consideration. This process resulted in the formulation of a common set which served as the primary input to the final recommendations presented in the Commissioner's report.

Findings Related to Testing, Assessment, and Evaluation

In contrast to *Let's Talk About Schools*, numerical summaries were not reported in the Royal Commission report. Instead, summary statements were made, supported by anecdotal information gleaned from the submissions made to the Commission.

With respect to testing, assessment, and evaluation, the Grade 12 Provincial Examination Program attracted the greatest attention of the general public, educational professionals, and students, similar to the situation for *Let's Talk About Schools* (see Table 1). Written by all grade 12 students who wish to graduate, these examinations were reintroduced in British Columbia in 1983, after having been discarded in the early 1970s. The decision to reintroduce provincial examinations was essentially a political decision. Three reasons were advanced by the then Minister of Education:

Firstly, our government believes strongly that students graduating from our high schools should be achieving challenging standards that are consistent across the province.... Secondly, a graduate applying for admission to a college or university should be assured of being treated equitably in the admissions process and be assured that high school marks will be taken at face value.... Finally, we introduce testing to reassure the public that we are serious about quality in our schools. (Henrich, 1983, p. 59)

No pressure was exerted from within the system to reintroduce Grade 12 "government" examinations. Indeed, many teachers and other educators at the time expressed opposition to their reintroduction (see Heinrich, 1983). That resistance has gradually waned; despite the controversy surrounding the reintroduction of these examinations, "the majority of students, parents, teachers of Grade 12 subjects, and the public wish them to be continued" (Marx & Grieve, 1988, p. 103). One particularly enthusiastic supporter said that "provincial exams were one of the best things to happen to B.C."

Not everyone, though, was supportive of the examination program. One student commented that, "The present system for provincially administered final exams is an unrealistic measure of the student's true skill and is valued far too highly in terms of the student's final mark." An anonymous brief stated categorically: "Get rid of government exams. They're destroying teaching programs and have too much weight."

One recurring issue with respect to the grade 12 examinations raised by supporters as well as opponents of the program concerned the weight given to the government exam in the final grade for a course. A number of recommendations were made to change the current division—50% teacher mark, 50% government exam—to accord more weight to the school mark. The most common recommended relative weighting was 60-40. Interestingly, during the time of the Royal Commission, the Student Assessment Branch, the division within the Ministry of Education responsible for the Grade 12 ex-

amination program, was reexamining the weighting issue in light of the *Let's Talk About Schools* finding that the majority of respondents felt that the 50% weighting given to the examinations was excessive. The Branch concluded its analysis with the recommendation that the current weighting be changed to 60-40. The same suggestion was made by the Commission's curriculum research team (Robitaille, Oberg, Overgaard, & McBurney, 1988, p. 49). However, following its analysis of learning and teaching functions and activities, the Commission's research team responsible for the learners argued that "the learners of British Columbia would be better served if the Provincial Exams were eliminated" (Marx & Grieve, 1988, p. 103) and replaced with an "emphasis on classroom practices that use criterion-referenced evaluation" (p. 106).

Following discussion leading to an integrated set of suggestions, the Commission recommended:

That provincial Grade 12 examinations be extended to all subject areas and that those examinations count for one-third of the student's Grade 12 marks. (Sullivan, 1988, p. 108)

Few comments were received by the Commission concerning the Learning Assessment Program and the province's participation in international studies. It seems respondents were more aware of and prepared to comment on issues related to "high stakes" examination programs which yield scores for individual students. They appeared to be either less aware of, or perhaps less concerned with, "lower stakes" programs which yield aggregated group scores. However, the curriculum research team (Robitaille et al., 1988, pp. 24-25) suggested that the purpose and procedures of the assessment program be more clearly tied to curriculum revision, and that sampling and testing techniques be employed in order to obtain system-level data on students' performance in the basic subject areas on an annual basis (p. 48). The Commission accepted these suggestions and recommended:

That the goals of the Provincial Learning Assessment Program be reformulated to concentrate on the provision of annual data on student achievements (knowledge, abilities, attitudes) in each of the basic subject areas. (Sullivan, 1988, p. 111)

Summary

Taken as a whole, the data and information obtained in the *Let's Talk About Schools* study and by the Royal Commission on Education suggest that there is far more support for than opposition to the policy directions announced by the Minister of Education on January 27 toward new and expanded examination, assessment, and evaluation programs. More than 90% of public and professional respondents to the *Let's Talk About Schools* discussion held that standards should be set for all forms of schooling, including home schooling and schools that operate outside the public system. Eighty-two percent of the public and 67% of the educators who expressed their views indicated that standards should apply to all grades. Further, public and professional endorsement of testing climbed steadily grade by grade, with more than 90% of respondents supporting the use of Grade 12 provincial examinations (Table 1, Item 2.4). There is clearly a belief among both the public and professionals

that standards and testing are useful ways to monitor and evaluate school effectiveness and, thereby, to ensure that quality exists within the system.

Further evidence of support for the Minister's policy directions can be inferred from the reactions to the Royal Commission. Concurrent with the release of the Commission's report, the Minister of Education appointed an Education Policy and Advisory committee composed of representatives of the major education professional organizations and community organizations with interests in education. The purpose of this Committee was to advise the Minister regarding the implementation of the Commission's recommendations. To assess further reactions to the Commission's recommendations, the Minister met with student, parent, teacher, principal, and trustee groups in 35 school districts located throughout the province. The consultation process demonstrated that "there is strong support for the great majority of the Commission's recommendations" (Brummet, 1989, p. 7). Of particular relevance here is the observation that with the exception of the Commission's recommendation that Grade 12 examinations count for one third of a student's Grade 12 grade, all other recommendations related to testing, assessment, and evaluation were accepted.

Despite this strong support for testing and assessment in British Columbia, the testing community in this province as elsewhere (see Madaus, 1985), must continually monitor the way in which tests are developed and used. We should remain mindful of the concerns raised by those who question the "overuse" of tests. For example, we saw two of the advisory research teams to the Royal Commission disagree on the continued use of Grade 12 government examinations. This apparent contradiction is centered on the widespread concerns that the content of the examinations has unduly influenced the curriculum and that teachers teach only what is considered likely to be tested. In the same vein, tests are frequently criticized for apparently focusing only on that which is easy to test: recall of factual information, application of routine algorithmic procedures, and the like. We need to be aware of such concerns. We must be prepared to give the consumers the protection they deserve by being mindful of testing, assessment, and evaluation as practiced and to reassert the fallibility of measurement even in the face of those who are obsessed with the infallibility of numbers.³

Notes

1. While the majority of educators believe that standards and testing are useful in monitoring school effectiveness and quality, they are concerned about testing practices which may impact negatively on students' academic growth and self regard.
2. The six areas were Schools and Society, Learners, Curriculum, Teachers, Finance, and Governance and Administration.
3. Toward this end, the British Columbia Ministry of Education commissioned in March 1989 a study team to assess by March 1990 the impact of the Grade 12 program on British Columbia's educational system.

References

- Brummett, A.J. (1989). *Policy directions: A response to the Sullivan Royal Commission on Education by the Government of British Columbia*. Victoria, BC: Ministry of Education.
- Heinrich, J. (1983). Why provincial examinations? *The B.C. Teacher*, 63(2), 59-60.

- Madaus, G.F. (1985). Public policy and the testing profession—You’ve never had it so good? *Educational measurement: Issues and practice*, 4, 5-11.
- Marx, R., & Grieve, T. (1988). *The learners of British Columbia*. Victoria, BC: Ministry of Education.
- Ministry of Education. (1985). *Let’s talk about schools: A report to the Minister of Education and the people of British Columbia*. Victoria, BC: Author.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: Department of Education.
- Petrie, H.G. (1987). Introduction to “evaluation and testing.” *Educational Policy*, 1, 175-180.
- Robitaille, D.F., Oberg, A.A., Overgaard, V.J., & McBurney, J.T. (1988). *Curriculum in the schools of British Columbia*. Victoria, BC: Ministry of Education.
- Sullivan, B.M. (1988). *A legacy for learners: The report of the Royal Commission on Education*. Victoria, BC: Ministry of Education.

DAN G. BACHOR

University of Victoria

Toward Improving Assessment of Students with Special Needs: Expanding the Data Base to Include Classroom Performance

In this paper, two traditional methods of assessment (the medical model and the psycho-educational model) are reviewed with the objective of illustrating the need to expand the data base included in assessment procedures. It is argued that there are at least two errors of measurement associated with traditional assessment practices: error due to the focus on a single measure or group of measures and error due to the restricted sample usually associated with these assessment practices. Following this review, a model of authentic measurement is introduced. In implementing this assessment model, a stronger emphasis is given to collecting descriptive classroom-based information over time and to evaluating all obtained data more critically in terms of what youngsters are required to do during assessment. Finally, it is argued that in making these changes, more effective instructional decisions can be made for students with special needs.

From the perspective of the special educator, measuring the classroom performance of normally achieving students is relatively straightforward. Such students are fairly consistent in their classroom performance. Thus any estimate of classroom achievement is a reasonable approximation of what has been learned to date. Further, when faced with some form of classroom assessment the majority of normally achieving students do not get upset or refuse to cooperate. They tend not to have a lengthy history of negative experience with assessment that must be overcome. In addition, normally achieving students typically do not have difficulty reading or listening to instructions. Such students usually possess sufficient skills so that (a) they know how to prepare for most classroom tests, (b) they can prepare at least a general test-time utilization plan, and (c) they are capable of making sense out of test items. Students with special needs often lack such test preparation and completion skills. (The term *students with special needs* will be used in the majority of cases throughout the remainder of this article for the sake of brevity. It should be remembered, however, that the term *special needs* normally is used to describe a wide range of students but is restricted here to students with mild to severe learning and/or emotional difficulties.)

Obtaining valid measures of classroom performance is equally important for students with special needs as it is for normally achieving students. Two main problems remain to be overcome before the abilities of students with special needs can be described fairly and usefully. The first is how to incorporate classroom data into existent diagnostic-prescriptive models. The

Dan Bachor is with the Department of Psychological Foundations, Faculty of Education. He is interested in assessment, learning strategies, and youngsters with learning difficulties.

second is how to obtain technically adequate classroom-based diagnostic information.

A few assessment techniques, appropriate for either classroom or clinical-diagnostic use, are described. The advantage of these techniques is that they can be used to assess students who are achieving normally or students who have special needs. In addition, the use of the suggested assessment techniques is incorporated into an educational decision making model. The procedures discussed have been derived from over 15 years of classroom and clinical experience with youngsters who have displayed learning and/or emotional handicaps. Before explaining this method, I review clinical-diagnostic assessment practices critically and give a rationale for changing assessment practices for students with special needs.

Current Clinical-Diagnostic Assessment and Instructional Practice

Psychoeducational assessment currently is the most widely practiced model of clinical-diagnostic or diagnostic-prescriptive assessment. It has emerged as the dominant model of assessment practice in a variety of facilities, ranging from neighborhood schools to specialized clinics and units of children's hospitals (e.g., Forness et al., 1980; Sattler, 1988; Ysseldyke et al., 1983). Although the model is implemented in slightly different manners as a function of the youngsters being assessed, most typically a school psychologist administers and scores a series of individual standardized tests. Interpretations of the test performance of students with special needs are then based on test score patterns or discrepancies between test and subtest scores (Hessler, 1986; Reynold, 1985).

When proposed, this approach was a departure from the predominant medical model. In the latter strategy, as might be inferred from the name, the purpose was to isolate causal explanations of poor performance, which were often assumed to be specific to an individual's constitutional makeup. Advocates of the psychoeducational model called for a shift from causal explanations to correlational hypotheses. The new assessment goal is to define a series of factors (related to how youngsters learn and behave) that contribute to lack of success in school.

As the psychoeducational model has been implemented, however, educators have come to depend on other individuals to obtain much of the information about students with special needs. Most frequently psychologists are requested to administer an assessment battery and to prepare a report in which students' learning patterns or personality are described. Extensive use has been made of norm-referenced measurement (Morrison, 1984; Smith, 1982; Stone, Cundick, & Swanson, 1988). Measures administered have included individually administered intelligence, neuro-psychological, and behavior-personality tests. Depending on the reason the youngster has been referred for assessment, other specialists may be involved to check, for example, degree of visual impairment or hearing loss. The primary objective is first to classify youngsters into special education categories, such as mentally handicapped or learning disabled, and then to determine an appropriate

education placement (Forness & Cantwell, 1982; Garrett & Brazil, 1979). Instructional planning follows after the student has been assigned to a classroom.

Accompanying the change from a medical model of assessment to a psychoeducational one have been shifts in the special education service delivery system. One of the most significant was the addition of a generalist school-based special educator: the learning resource teacher (Dunn, 1973; Hardy, 1977). The introduction of an assessment-instructional specialist in the home school has had ramifications for the assessment process. Learning resource teachers now have the role of assisting youngsters in the regular classroom. When first introduced, they typically started with a referral from the regular classroom teacher. This consisted of descriptive information about the student and his or her present problem or problems. The detail and precision of such obtained referrals were a function of at least two factors: (a) the classroom teachers' time to and skill in preparing referral information and (b) the type of information on which the teachers were basing their referrals. The learning resource teachers then would assess each youngster. Subsequently, instruction would be provided to small groups of students with similar difficulties, the students being removed temporarily from the regular classroom during this instructional period.

Thus a modification emerged in the practice of the psychoeducational model of assessment. The resource teacher assumed command of educational assessments, and the psychologist retained responsibility for the psychoeducational assessment used to determine program placement. In addition, a new layer of service delivery resulted, in which the learning resource teacher worked in cooperation with the classroom teacher to provide school-based instruction.

Challenges to the Psychoeducational Model

Although the application of the psychoeducational assessment model with students with special needs has been an improvement over the causal attributions resulting from the use of the medical model, its use now is being reexamined (Algozzine & Korinek, 1985; Gelzheiser, 1987). Two challenges have emerged recently. First, a challenge to the service delivery system and to the categorical system of classifying youngsters is emerging. Second, the adequacy of present assessment practices is being queried. Taken together, the implication inherent in these challenges to the current psychoeducational model is that several aspects of the present system of special education must be rethought.

Challenging the Categorically-Based Psychoeducational Service Delivery System

At the present time a challenge is emerging to the special education service delivery system (Hallahan, Keller, McKinney, Lloyd, & Bryan, 1988; Kaufman, Gerber, & Semmel, 1988; Keogh, 1988; Singer & Butler, 1987; Stainback & Stainback, 1984). In its most extreme form, proponents of what has been termed the "regular education initiative" are calling for dramatic changes such as the appointment of school-based personnel to facilitate the

integration of all handicapped individuals, the expansion of regular curriculum (including such topics as community living skills, sign language, and braille), and the retraining of special education teachers to focus on instructional categories rather than on student categories (that is, a noncategorical model). Once the extreme positions are tempered, however, there are two main objections to the psychoeducational model: (a) the categories presently used to classify students with special needs are neither valid nor reliable, and (b) the assessment process is costly, time consuming, and lacking in educational relevance.

Further, the relationship between classroom assessment and instructional procedures and those completed in the resource room have become remote. Once the referral has been made and the student accepted for special services, the classroom teacher and the special educator to a large degree work independently of each other. In part due to this independence, a significant discrepancy has emerged in the tools used to assess youngsters referred for special education and those assessment strategies used to determine their classroom performance. Although classroom observation has been utilized, classroom performance has not been used extensively in conducting psychoeducational assessment. As a result, placement decisions were not based on what students were capable of doing in the classroom from which they were referred. In addition, the learning resource teacher or other special education teachers have assumed a large portion of the responsibility for teaching the student with special needs, and the regular classroom teacher has turned his or her attention to the remainder of the class, completely distancing any student with special needs from his or her classmates. Thus the special educator has learned to rely on assessment and teaching strategies that are divorced from youngsters' daily classroom performance and expectation.

Challenging the Conception of Students

Measuring what students with special needs have learned is confounded by the ineffective learning strategies they typically display. Examples are searching for easy solutions, failing to persist at work, being poorly organized, setting unrealistic goals, and attacking work in a rigid manner. Having a cluster of faulty learning habits, however, is not the only significant marker. Equally important is the social value attached to these faulty learning habits; that is, an answer to the question, "Is the absence of specific learning procedures or of particular knowledge seen as a significant weakness?" must be obtained. If the risk associated with lacking some ability is minimal, there will be few negative consequences assigned to that incompetence. For example, the student who cannot sing may be embarrassed in music class, but the probability of being identified as having learning problems because he or she is a poor singer approaches zero. On the other hand, if the ability in question is reading, the chance of being so identified is high. The final aspect to be considered is hinted at in the last example; that is, how frequently will any ability, skill, or knowledge be required?

Placed in perspective, it seems logical to think of all youngsters as falling along a continuum of competence, rather than as being divided into a set of discrete categories. There will be some school specific tasks at which students will be successful and some at which they will need extra assistance in order to learn more effectively. The goal of assessors completing either classroom or clinical assessments, then, is to obtain a description of where students fall along this continuum. As discussed above there are three facets to be considered: (a) the degree to which students' learning strategies are faulty or to which they lack necessary knowledge, (b) the importance or social value attached to having either the appropriate learning patterns or information, and (c) the frequency with which such abilities, skills, or knowledge are required in the classroom. The greater the magnitude of and the more the interplay between these facets, the greater the perceived weakness and the stronger the negative valence attached to the description of the student: In short, the performance of youngsters must be viewed in a relative rather than in an absolute manner.

Challenging the Assessment Model

As they have begun to assume a more comprehensive assessment and instructional role within the neighborhood school, learning resource teachers slowly have started to question the adequacy of current assessment techniques. Two problems have been identified.

The first problem is that the process of making placement decisions has become cumbersome (Ysseldyke, 1983; Ysseldyke et al., 1983). To review, a series of tests is administered by both the school psychologist and the resource teacher. Other professional personnel are consulted on a case-by-case basis. After this process is completed, a school-based team (consisting of the school psychologist, counselor, principal, learning resource teacher, and referring teacher) meets to determine the disposition of the case. In evaluating this system of placing students, Ysseldyke (1983) reported a number of startling findings: (a) the topics of such meetings were dominated by reports given by school psychologists and the major focus was on achieved test scores; (b) the greater the number of tests reported, the more likely they were technically inadequate; (c) the issue of placement decisions was allocated the second largest percentage of meeting time and educational planning received the least attention; and (d) teachers and parents rarely were asked to offer information or to propose educational options.

A second common complaint has been with the usefulness of some of the commercially available tests, even though teachers have reported standardized measures to be the most useful in instructional planning (Thurlow & Ysseldyke, 1982). There are at least four major reasons for the restricted applicability of much of the currently available assessment data:

1. The technical adequacy of many measures is questionable.
2. The emphasis has been on reporting test scores rather than on descriptions of what students typically do.

3. In many tests there is an insufficient number of items per ability or skill being measured to allow a reasonable estimate of any set of students current level.
4. The conditions under which these data have been collected do not represent classroom performance.

Proposed Solutions Within the Current Model

Some solutions have been proposed within the application of the psycho-educational model in the attempt to better characterize both what and how students with special needs have learned. The first approach is to reexamine data collection technology and tests. The second strategy is to change testing procedures.

Expert Systems and Broadening the Test Base

The first of these proposed solutions has been to generate expert assessment batteries. The rationale has been to identify measures that would yield as close to perfect error-free diagnostic-prescriptive data as possible. In the attempt to devise such an expert system, computer-based automated assessment systems (e.g., Colburn & McLeod, 1983; Hofmeister & Lubke, 1986) and other specialized assessment batteries (e.g., Kaufman & Kaufman, 1983) have emerged. The special educator's role in assessment is again minimized with this proposed solution to assessment administration. Scoring and interpretation follow from using the expert system and often is computer generated.

An alternative approach has been to prepare behavior-specific or subject-based assessment instruments. A plethora of such measures, consisting of checklists, standardized tests, and nonstandardized published or unpublished informal inventories, have been developed (for a discussion see Salvia & Ysseldyke, 1988). The task of learning assistance (resource) teachers or other school-based special education teachers is much more complicated if they opt for this solution. They must select appropriate instruments that when administered and scored will provide useful instructional planning information. Wilson (1985) has argued that the process can be simplified considerably by using cut-off scores derived from administering group achievement tests as the base for placement decisions. He found that when using the 10th percentile as a cut-off score, approximately 70% of students with special needs were classified correctly. The use of such a simplified technique, however, would still mean nearly 30% were misclassified.

Changing the Data Collection Procedure

As mentioned earlier the most frequently employed assessment technique is the administration of standardized tests. Data collection procedures are being reexamined within the field of special education with the intention of obtaining a better representation of what and how students with special needs learn.

One alternative assessment strategy has been to utilize criterion-referenced measurement to sample student in-class performance. Two criticisms have been stated. First, a criterion-referenced approach to measurement has

the implicit assumption that the subskills of any curriculum can be stated and items defined to measure such subskills. Second, to fully represent any curriculum an inordinate number of items is required; however, in contradiction, a parallel criticism is that most criterion-referenced tests contain too few items per objective to provide useful diagnostic information (Bachor, 1985). One answer has been to refine the procedures used in criterion-referenced measurement. In its new form in special education, criterion-referenced measurement has been termed curriculum based assessment or CBA (Deno, 1985; Fuchs, Fuchs, Benowitz, & Barringer, 1987; Marston, Mirkin, & Deno, 1984). CBAs are measurement procedures that are designed to match students' program objectives.

There are two main versions of CBA. In the version suggested by Fuchs and her colleagues (Fuchs, Deno, & Mirkin, 1984; Fuchs & Fuchs, 1986; Fuchs et al., 1987), a pool of domain-referenced items is constructed to match annual goals and instructional objectives. Items, called probes, are then sampled from the item pool and administered to students to monitor instructional effectiveness. In the version suggested by Deno and his associates (Deno, 1985; Deno, Marston, & Mirkin, 1982; Wesson, Fuchs, Tindal, Mirkin, & Deno, 1986), quick probes to estimate performance, for example, the number of words read correctly in one minute from a basal reader, the number of correctly spelled words from dictated lists, and the number of words written on a story starter, are obtained.

Wesson et al. (1986) summarized the use of CBA by stating that using it allows a teacher to develop short, simple measures based on curriculum. These measures are administered from three to five times a week and students' performance data are graphed and analyzed. This information then is used to determine the appropriateness of the students' individualized educational programs. Fuchs and Fuchs (1986) argue that the use of CBA in developing and evaluating instructional programs over the long term can be expected to have a significant positive effect on the achievement of students with special needs. They add that for the use of CBA to be effective, instructional goals for individual students must be stated and monitored. In a meta-analysis in which the goal was to compare long-term and short-term assessment, they noted an effect size interaction. "When progress was measured toward long-term goals, effects sizes calculated on global outcome measures were higher than on probe-like outcomes. On the other hand, when progress was measured toward a series of short-term goals, effect sizes were lower on global than on probe-like outcome measures" (Fuchs & Fuchs, 1986, p. 78).

Summary

While such revisions may be helpful to anyone concerned with diagnostic procedures, they must be put into perspective. All collected data must be seen to contain error in two senses:

1. There is the error of measurement associated with the administration of any particular tool in the sense that no single procedure, regardless of the attractiveness of the assessment technique, can stand alone.

2. There is the error of the measurement that accompanies taking time-bound samples, in that in taking any individual measurement only what a person did at that time can be estimated. In the decision making model to be addressed next, these two problems areas are addressed and techniques to expand data collection are suggested.

*Modifying the Model: Incorporating Decision Making
and Classroom Data*

A second assessment model—coined the decision making model (Bachor & Crealock, 1986), a decision making procedure (Ysseldyke & Regan, 1980), or an instructional-based model (Humphries & Wilson, 1986)—has been suggested in which the traditional model of information collection has been modified and extended. The primary emphasis in implementing the decision making model is to obtain instructionally relevant information about both students and teaching techniques. Because this model is an extension of the traditional one, several hybrids of each are currently being practiced by special educators (e.g., Ministry of Education, 1989).

Some Suggested Assessment Strategies

Assessment when using this model changes to a multifaceted process designed to reflect a continuum of strengths and weaknesses. The argument is that no one assessment technique will suffice. Stated more positively, the application of all assessment techniques, regardless of validity, can never result in more than a situation-specific estimate of what a youngster can do. Thus, to obtain a reasonable estimate of performance for any youngster, a number of assessment techniques (interviewing, observing, testing) must be employed that yield information about performance under a variety of conditions (such as when receiving instruction, when working independently, when doing something enjoyed, etc.). By following such procedures, information will be obtained that will allow an estimate of the degree of difficulty experienced by any individual in doing something. Or the use of such procedures will allow a special educator to approximate where any student with special needs falls along the learning/achievement continuum. In addition, it is recognized that a portion of the collected data must be drawn from the youngsters' home classroom. The procedures of cumulative record analyses, comparative and continuous observation, and work sample analysis are suggested for this purpose. These procedures are accompanied by conducting interviews and examining the effectiveness of instructional procedures.

Starting with the students' history. In applying the decision making model, Bachor and Crealock (1986) argue that the special educator should start by examining the track record of students with special needs. They recommend classifying all available information as being one of four types: historical opinions, current opinions, historical measurements, current measurements. In a recent revision, Bachor (1989) suggested that teachers should further classify the information in the cumulative files into crude commonalities by describing the conditions under which the data were collected. In completing this procedure they are to describe what the assessor did (interviewing, observing, testing, teaching) while presenting any assess-

ment task and what the youngster was required to do in response (listen, write, read, copy, talk, make). In addition, any available information on how the student has been taught and what subject-specific materials have been used should be noted, especially if there are data to note the utility of such methods and materials. Next, the data are to be grouped (listing the type and sources of the collected data) around any probes or tentative hypotheses that the teacher wishes to advance as the starting point of further assessment. (A probe is a data point that only occurs once but one that the decision maker suspects should be checked further. A hypothesis differs from a probe in that it must be supported by two or more data points from different sources, preferably those that cover a span of time rather than those that were conducted on the same day.)

Sampling classroom performance by observation. As indicated in the introduction to this article, when students with special needs are given a classroom test they face a different set of conditions than do students who are normally achieving. Many of the currently applied classroom and clinical assessment procedures are designed so that the obtained data do not reflect the knowledge-base of students with special needs. Comparative observation is a technique in which teachers examine students' classroom performance in any subject matter. They do so by comparing the performance of the student or students in question to a stratified sample of other students in the class. To be most useful this sample should be subject matter specific and should consist of two to four students who fall into three groupings: those performing near the top of the class, those performing in the average range, and those having difficulty. Evaluation standards then are based on the reason for referral; they can range from amount of work completed in a given period to the quality of the completed assignments.

This comparative evaluation may be accompanied by a continuous chronology. The purpose of obtaining such a chronology is to provide a descriptive summary of the student's activities in his or her classroom environment. The subject matter observed is not critical in this case; however, the conducted observation should include as wide a variety of activities as is possible across the school day, while avoiding atypical events such as testing periods, in order to obtain a complete representation of any student's general pattern of behavior.

Work sample analysis. Bachor (1979) has described a procedure whereby students' daily classroom performance is used as the basis of assessment. This procedure has been revised to allow the regular classroom teacher or the special education teacher to obtain a technically adequate sample of classroom performance that can be translated directly into instructional objectives (Bachor, 1984, 1989). Specifically, the following four steps are suggested to complete work sample analysis when the purpose is the description of a single student or a small group of students:

1. Select three to five assignments completed over a period of two to six weeks. The skills or abilities to be observed must be repeated in some form in each selected assignment. For each assignment, list strengths (abilities or skills at or above grade or age level expectations), skills

present (abilities or skills mastered by individual but not meeting grade or age level standards), and weaknesses (errors of omission or commission).

2. Look for commonalities in error patterns across all work sample analyses.
3. Compare observations of strengths, skills present, and weaknesses to expectations set for students (use curriculum guides or other locally established standards).
4. Using both commonalities and expectations, prepare diagnostic test items to establish approximate boundaries of performance.

Procedural interviews. Estimates of how students are learning may be obtained by interviewing them (Meyers, Pfeffer, & Erlbaum, 1985). Interviews can be conducted in three separate phases. First, students can be asked to explain how they plan to do any task. Second, they can be interviewed while completing a preselected standardized assignment similar to classroom-based activities. Third, they can be questioned for post hoc explanations of how something was done. Again comparative interviews can be conducted to determine, in a normative yet descriptive manner, where one student stands in relation to another.

There are difficulties associated with interviewing, however. If one administers a series of subject-specific problems with the intention of establishing procedural knowledge, the problem is that to some lesser or greater extent such performance must be inferred. A wide discrepancy in subjects' reading levels, for example, will be observed as variability in text processing and likely as restricted word knowledge (Kintsch, 1988). Thus, at a minimum, steps should be taken to observe subjects' reading levels and to describe their reading strategies across a variety of texts. Similarly, students' abilities to remember material may influence their strategy knowledge and utilization, as will their confidence in approaching school-related subjects. For example, Steacy & Bachor (1987, 1988) found that functional short-term memory span and academic self-concept affected students' acquisition and performance on arithmetic word problems. Thus it appears necessary to interview youngsters on strategies across a variety of subject matters to note variations in their strategic repertoire and their affective responses to how well they are doing.

Effectiveness of instruction. A final assessment procedure within this model is to conduct trial lessons. The purpose here is to evaluate whether any suggested instructional method or material is likely to work well with a group of students. The criterion of success is whether the student can demonstrate mastery of the taught curriculum at the end of the instructional period. Given that these procedures are intended for students with special needs, a further precaution is taken; that is, can such students remember the instruction, and did they remember it over time?

Summary

The key difference in implementing a decision making model, as opposed to the psychoeducational model, is that the emphasis is on description of perfor-

mance. To translate diagnostic procedures into instruction, assessment procedures need to be drawn from a variety of sources and should reflect how students act under different conditions. Making such modifications in the assessment process should enable all professionals to better relate assessment to required instruction.

Conclusion

The position adopted in this article is that yet another directional shift in the assessment and instructional practices for students with special needs must be taken from a psychoeducational model of assessment to an ecological one. This modification in procedure is fairly straightforward. Instead of relying on single measurements or time-bound measurements, measurement portfolios that represent any youngster's classroom ecology are required. In building such portfolios, the emphasis will need to be on information utility. The question of utility can be stated as follows: How do the data being collected help to make more effective program and placement decisions? To answer this question, it was suggested that descriptive data that span time and contain both objective and subjective information are to be collected. Further, it was recommended that such descriptive data should be gathered on a stratified sample of students to permit portfolio comparison. By collecting these assessment portfolios and using the obtained information to monitor instructional methods and materials, more effective instructional decisions can be made for students with special needs.

References

- Algozzine, B., & Korinek, L. (1985). Where is special education for students with high prevalence handicaps going? *Exceptional Children*, 51, 388-394.
- Bachor, D. (1979). Using work samples as diagnostic information. *Learning Disabilities Quarterly*, 2, 45-52.
- Bachor, D. (1984, February). *Children's work: A basis for diagnosis and remediation*. Paper presented at the North Central Teachers' Association Conference, Edmonton, AB.
- Bachor, D. (1985). Some trends in assessment and curriculum and instruction for mathematics in special education: An international perspective. *Focus on Learning Problems in Mathematics*, 9, 175-180.
- Bachor, D. (1989, January). *Effective assessment and instructional planning for the learning disabled*. Paper presented at the Manitoba Congress of the Council for Exceptional Children Conference, Winnipeg, MB.
- Bachor, D., & Crealock, C. (1986). *Instructional strategies for students with special needs*. Scarborough, ON: Prentice-Hall.
- Colburn, M., & McLeod, J. (1983). Computer guided educational diagnosis: A prototype expert system. *Journal of Special Education Technology*, 6(1), 30-39.
- Deno, S.L. (1985). Curriculum-based assessment: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Deno, S.L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children*, 48, 368-376.
- Dunn, L.M. (1973). *Exceptional children in the schools: Special education in transition*. New York: Holt, Rinehart and Winston.
- Forness, S.R., & Cantwell, D.P. (1982). DSM III Psychiatric diagnoses and special education categories. *Journal of Special Education*, 16, 49-63.
- Forness, S.R., Urbano, R., Rotberg, J., Bender, J., Gardner, T.P., Lynch, E.W., & Zemanek, D.H. (1980). Identifying children with school learning and behavior problems served by interdisciplinary clinics and hospitals. *Child Psychiatry and Human Development*, 11(2), 67-78.

- Fuchs, D., Fuchs, L.S., Benowitz, S., & Barringer, K. (1987). Norm-referenced tests: Are they valid for use with handicapped students? *Exceptional Children*, 54, 263-271.
- Fuchs, L.S., Deno, S., & Mirkin, P. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, 21, 449-460.
- Fuchs, L.S., & Fuchs, D. (1986). Curriculum-based assessment of progress towards long-term and short-term goals. *Journal of Special Education*, 20, 69-82.
- Garrett, J.E., & Brazil, N. (1979). Categories used for identification and education of exceptional children. *Exceptional Children*, 45, 291-292.
- Gelzheiser, L.M. (1987). Reducing the number of students identified as learning disabled: A question of practice, philosophy, or policy? *Exceptional Children*, 54, 145-150.
- Hallahan, D.P., Keller, C.E., McKinney, J.D., Lloyd, J.W., & Bryan, T. (1988). Examining the research base of the regular education initiative: Efficacy studies and the adaptive learning environments model. *Journal of Learning Disabilities*, 21, 29-35.
- Hardy, M. (1977, Spring). The learning resource teacher. *Education Canada*, 17, 28-32.
- Hessler, G.L. (1986). The issue of intellectual ability in severe discrepancy analysis for the determination of learning disabilities. *Remedial and Special Education*, 7(4), 58-61.
- Hofmeister, A.M., & Lubke, M.M. (1986). Expert systems: Implications for diagnosis and treatment of learning disabilities. *Learning Disability Quarterly*, 9, 133-137.
- Humphries, T.W., & Wilson, A.K. (1986). An instructional-based model for assessing learning disabilities. *Canadian Journal of Special Education*, 2, 55-66.
- Kauffman, J.M., Gerber, M.M., & Semmel, M.I. (1988). Arguable assumptions underlying the regular education initiative. *Journal of Learning Disabilities*, 21, 6-11.
- Kaufman, A.S., & Kaufman, N.L. (1983). *K-ABC: Kaufman assessment battery for children*. Circle Pines, MN: American Guidance Service.
- Keogh, B.K. (1988). Improving services for problem learners: Rethinking and restructuring. *Journal of Learning Disabilities*, 21, 19-22.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.
- Marston, D., Mirkin, P., & Deno, S. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *Journal of Special Education*, 18, 109-117.
- Meyers, J., Pfeffer, J., & Erlbaum, V. (1985). Process assessment: A model for broadening assessment. *Journal of Special Education*, 19, 73-89.
- Ministry of Education. (1989). *Special education manual of policies and procedures*. Victoria, BC: Province of British Columbia.
- Morrison, H.E. (1984). Using difference scores in making screening and placement decisions. *Diagnostique*, 9, 178-184.
- Reynold, C. (1985). Measuring the aptitude-achievement discrepancy in learning disability diagnosis. *Remedial and Special Education*, 6(5), 37-48.
- Salvia, J., & Ysseldyke, J. (1988). *Assessment in special and remedial education* (4th ed.). Boston, MA: Houghton Mifflin.
- Sattler, J. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Sattler.
- Singer, J.D., & Butler, J.A. (1987). The Education for All Handicapped Children Act: Schools as agents of social reform. *Harvard Educational Review*, 57, 125-152.
- Smith, J. (1982). The Kaufman assessment battery for children: A major advance in intelligence testing for educational intervention. *Counterpoint*, 3(2), 1, 22.
- Steady, N., & Bachor, D.G. (1987, June). *Learner characteristics and task variables as predictors of rule induction ability on word problems*. Paper presented at the Canadian Society for Studies in Education Conference, Hamilton, ON.
- Steady, N., & Bachor, D.G. (1988, June). *A description of selected grade 6 rule induction behavior when solving arithmetic word problems*. Paper presented at the Canadian Society for Studies in Education Conference, Windsor, ON.
- Stainback, W., & Stainback, S. (1984). A rationale for the merger of special and regular education. *Exceptional Children*, 51, 102-111.
- Stone, B., Cundick, B. P., & Swanson, D. (1988). Special education screening system: Group achievement test. *Exceptional Children*, 55, 71-75.

- Thurlow, M.L., & Ysseldyke, J. (1982). Instructional planning: Information collected by school psychologists vs. information considered useful by teachers. *Journal of School Psychology, 20*, 3-10.
- Wesson, C., Fuchs, L., Tindal, G., Mirkin, P., & Deno, S.L. (1986). Facilitating the efficiency of on-going curriculum-based measurement. *Teacher Education and Special Education, 9*, 166-173.
- Wilson, L.R. (1985). Large scale learning disability identification: The reprieve of a concept. *Exceptional Children, 52*, 44-51.
- Ysseldyke, J. (1983). Current practices in making psychoeducational decisions about learning disabled students. *Journal of Learning Disabilities, 16*, 226-233.
- Ysseldyke, J., & Regan, R.R. (1980). Nondiscriminatory assessment: A formative model. *Exceptional Children, 46*, 465-466.
- Ysseldyke, J., Thurlow, M., Graden, J., Wessen, C., Algozzine, B., & Deno, S. (1983). Generalizations from five years of research on assessment and decision making: The University of Minnesota Institute. *Exceptional Education Quarterly, 4*(1), 75-93.

LESLIE D. McLEAN

Ontario Institute for Studies in Education

Time to Replace the Classroom Test with Authentic Measurement

An argument is made that test scores are painfully feeble and fall far short of adequacy as measures of educational achievement. We have learned a great deal in the past decade about the importance of meaning and context in language teaching, and these ideas are beginning to inform our understanding about the learning of mathematics and science. Achievement takes on new meaning—not as some general or abstract trait to be captured by a single number but instead as performance on meaningful tasks in a recognizable context. The most promising tool for the assessment of authentic achievement in the classroom (and perhaps for accountability as well) is the portfolio, a systematic, cumulative record of performance. Evaluation of portfolios, or “records of achievement” as they are known in the United Kingdom, has shown them to be promising ways to truly integrate student evaluation with the rest of teaching and learning.

It has been fashionable to define *test* as any systematic measurement procedure, including observations, checklists, and other tools in the category, but that broad definition has never caught on. To the vast majority of people, a test is a fixed set of questions written on paper to which one responds and from which the test administrator derives a test score. Say “classroom testing” to anyone and they will imagine students “writing” a test. Other evaluation tools are used in classrooms, of course, but they are not referred to as tests. Measures of listening comprehension, for example, have spoken stimuli, but the responses are written and the outcome is a test score. Evaluation of spoken language is done through an oral interview—not a test. In this article, the word test will be used in the specific sense of written stimuli, written responses, and a single number (or a few numbers) as the outcome.

Bloom’s Taxonomy (Bloom, 1956) struck a responsive chord, but we have not seen the expected flowering of items testing the higher levels—synthesis, evaluation, and the like. In fact, every study ever done of actual tests has found a heavy, often exclusive, reliance on questions that ask only for knowledge. We have built a structure of so-called test theory that was attractive on the outside but has remained almost empty as far as achievement tests are concerned. It is the thesis of this article that the achievement test structure is mostly empty because it is uninhabitable. We have great difficulty seeing that test scores are a cop-out, an admission of ignorance we can no longer afford; as measures of educational achievement, they are painfully feeble and fall far short of adequacy. The persistent urging of tests on classroom teachers and the neglect by measurement specialists of any other tools continues to be a great disservice to education.

Leslie McLean is a professor and head of the Educational Evaluation Centre. His research interests center on student evaluation.

The pity is that the structure we call test theory and the instruments derived within it threaten to drive out any competitors, despite their vacuousness when it comes to the measurement of achievement in the classroom. They are vacuous because they are empty of the content that marks authentic achievement. A soul-searching article on measurement (i.e., test) driven instruction (Airasian, 1988) is illustrative, in that it raises many questions about content but provides no answers. While we are waiting, there are content-rich alternatives to tests—more positive and attractive tools for the evaluation of student achievement. The rest of this article is devoted to a description and discussion of an alternative class of tools and to an argument as to why they are more likely to be authentic and why they deserve to replace testing in the classroom.

Authenticity

The concept of validity has long been prominent in psychological testing, where nomological nets can be woven. For example, the Myers Briggs Type Indicator (MBTI) has a strong claim to validity within a Jungian theoretical framework. There has been a call for some years for a theory within which interest measures can be grounded; the early enthusiasm for empirical scales such as the Strong or Strong-Campbell is giving way to skepticism. In achievement tests, however, validity has proved an elusive quality. Credible achievement constructs (Haertel, 1985) have not been found. Face validity or curricular validity (essentially an overlap with the surface curriculum) has had to suffice, but given the complexity of curricula, the tests invariably had low validity. Large-scale surveys based on item and person sampling achieved much more complete overlap, but they tended to fall of their own weight back on multichoice items at the knowledge level. Before the miracle of item response models, such surveys did not yield scores for individuals and hence were limited to monitoring. They therefore made little or no contribution to measurement in the classroom beyond providing extremely rough guidelines and items that could improve only the poorest existing practice. We now have the Duplex Design claiming to provide individual scores (Bock & Mislevy, 1988), but not everyone is willing to accept the claim (Goldstein & McLean, 1987). Even if accepted, few jurisdictions will be willing or able to afford such large surveys on a regular basis, so impact on classroom measurement will be minimal.

Impact on classroom practice can be substantial in that high-stakes tests lead to more practice with test items in order to achieve better test scores (Corbett & Wilson, 1987). Improved test scores, however, do not lead to improved performance on meaningful tasks, as language teachers found a decade ago. Reduction of curriculum content to test content, common in science and mathematics classes, cannot be described as improvement. Validity is now being addressed by what we might call NAEP Norms, citing the items answered correctly by high proportions of the survey sample at various levels of the test score. This is an improvement on previous practice in that the interpretation is based on performance on items we can actually examine. When we examine them, however, we are thrown back on prior

questions, “Why are these the right items?” and “Why should we select items to examine based on this particular test score?” We need, in short, a way to authenticate our measures. Authenticity may be a more helpful and useful word than validity to describe the quality we seek in achievement measures.

An articulate call has recently gone out for authentic measures, endorsed and distributed by the United States National Association of Secondary School Principals (Archbald & Newmann, 1988). The concept of authenticity is based where it should be, on the theories and models of curriculum content and cognitive processes (where we have them), rather than on abstract models of test scores masquerading as theory. There are, for example, communicative or functional language theories that guide curriculum design, including the definition and measurement of achievement. There do not appear to be comparable theories in other subjects, mathematics and science for example, and it may be that there are no others. Perhaps there are only different versions of language theory, versions that apply to the different symbol systems used in mathematics and science—and music and art. Cognitive scientists are hard at work on theories of microprocesses and macrostrategies that might reach theory status.

An achievement measure is authentic if it faithfully reflects:

1. The knowledge and the processes that form the basis of the subject matter. (Cognitive scientists refer to executive processes, language theorists refer to meaning making and discourse competence. The concept of process is not exactly a firm one yet, but it appears over and over again in various guises.)
2. The conditions under which the achievement normally takes place.

Some years ago, a seminal study of teacher constructs of the curriculum and of students went beyond the surface curriculum (Bussis, Chittenden, & Amarel, 1976), and now our achievement measures are starting to follow (Swain, 1988; McLean, 1988).

Subject Matter Processes and Knowledge

The only subject matter processes and knowledge we understand at all well are in the subject of language (Canale & Swain, 1980; Gorman, 1986; Thornton, 1986). Three points are salient for this paper.

The ability to communicate (whether listening or reading, speaking or writing) depends crucially on what one already knows—on the knowledge, concepts, and beliefs acquired to date. Meaning does not exist in some pure form outside us; it is constructed in the process of communication from what we already know or believe. Knowledge of the world is therefore crucial to communication.

Discrete point knowledge (isolated words, rules of grammar) tells us nothing useful about what a person can do with it.

Achievement worth knowing about involves subtle changes that happen over months. Changes are continuous and cumulative, the result of multiple interactions of which schooling is but one. A necessary condition for change is the construction of meaning.

Conditions Under Which Achievement Takes Place

The conditions of learning and the conditions for the demonstration of achievement have not featured in the design of achievement measures to date. Once designed, published tests come with instructions for standard administration, but this is a surface feature different from the point being made here. Language theories emphasize the importance of context and purpose in the shaping of communication, so both pedagogy and measurement should take these into account. Contexts meaningful to the students are required for teaching, hence for achievement measurement. Communicative language achievement measures present texts of a hundred words or more set in a context with which the examinees can identify. It is therefore impossible to produce measuring instruments that can be used in the same way across a country, or even a province. The instruments (substantial booklets are required) can be used, but some parts of them are not appropriate and universal scoring schemes are not valid. They have to be regarded as sample collections of stimuli to be adapted to local conditions. (Examples are the reading assessment booklets used by the Assessment of Performance Unit in England and Wales and the procedure, *A vous la parole*, devised by the OISE Modern Language Centre for use in French immersion classes. The APU instruments were applied across England and Wales, but the reports were of disaggregated results and the most valued outcomes have been the booklets themselves which have been widely distributed to teachers.)

Students are increasingly being encouraged to work cooperatively in small groups. This is being promoted in language classes, and the umbrella administration for Toronto schools recently published a book encouraging cooperative small group learning across the curriculum (Metropolitan Toronto School Board, 1988). The OISE Cooperative Learning Centre is busy with professional development workshops, and there is an International Association for the Study of Cooperation in Education. If much of the classroom learning has taken place in groups, any measurement of achievement that relied exclusively on individual measurement would lack authenticity. It may be acceptable to require individual products of learning, but students should have the opportunity to develop some of them in cooperation with others. If one of the goals of the curriculum is for students to learn to work well in groups, then some assessment of group processes would appear to be required.

Summary of Implications

Authentic measurement must involve varied settings and purposes. It must include assessments collected over several months, the longer the better. Tasks meaningful to the students are required. All these conditions predict that performance will vary with tasks in complex ways. Student-task-setting interactions will mean that authentic measures have very low indices of internal consistency (Swain, 1988). Authenticity will be inversely related to traditional estimates of measurement quality, because the traditional measures are derived from testing constructs that do not accord with current

subject theories. This is not surprising, because test theory was assumed to be independent of subject theories.

The Forms of Authentic Measurement in Classrooms

Teachers and students in classrooms are uniquely situated to respond to the implications of subject theories, at least as we see them in language. The implications for achievement definition and measurement are profound. What is required is a systematic, cumulative record of performance on meaningful tasks. The writing folder can be one example of such a record, if the teacher sets the tasks systematically and well, in cooperation with the students. Standards are set by the requirements for the contents of the folder over the course of a term, or better over a school year. Evaluation of individual entries is done jointly by teacher and student, sometimes bringing in other students. Students often work cooperatively, but each student retains a collection of the work he or she produces. Not all work is marked, but the teachers specify in advance a set of assignments that are expected to be in the folder for marking. Students choose some of their best work to add to the record, negotiating a mark with the teacher. In case of dispute, both marks can be assigned (the student's and the teacher's) with a rationale. Teachers who have tried this report that the negotiations have been some of the most valuable learning experiences of the term (Radnor, 1988).

The generalized form of the systematic, cumulative record is sometimes called a portfolio, or just a record of achievement. Records of achievement have been the object of study in 10 large pilot schemes in England and Wales over the past two years, and the evaluation of those schemes has just been published (Broadfoot, James, McMeeking, Nuttall, & Stierer, 1988). Many of the schemes experimented with records of achievement in and out of school and in nonacademic areas, raising a number of ethical questions. Some observers are concerned, for example, that records of achievement that include activities outside of school (clubs, part-time jobs) will be used by the authorities for surveillance (Hargreaves, in press; *E+M Newsletter*, 1988-89).

Arriving at School Marks

Portfolios lend themselves best to category marks such as A, B, C, the difference from usual practice being that descriptions of the categories can be written in terms of the performance expected in the portfolio. Portfolio entries are marked individually and the marks aggregated to arrive at the term or final mark. The advantage of such a scheme is that the weighting of marks is explicit and visible to students and parents. Teachers in secondary schools often weight term tests, quizzes, and assignments, of course, but the weighting of content in the tests is rarely known accurately. Percentage marks are used, with a spurious appearance of accuracy and meaning. Once the complexity (especially variability) of achievement is better known through the subject theories, it is easier to accept that category marks (at most five) represent the most accurate assessment available to the teacher—hence a more honest procedure than the ones usually seen.

Reporting to Parents

Parents receive the usual report cards, with category marks, accompanied by definitions of the categories. Portfolios are collected by the teacher at the time of parent interviews so that conversations can be conducted in terms of the demonstrated performance of the student. (Parents may have seen the portfolio all along, of course, because it is in the student's possession most of the time.) There is bound to be some normative interpretation when the teacher justifies the standards set for the various marks, for example, "These assignments were completed successfully by all but three students in the class, but the extensions proved much more difficult. Your daughter did all the difficult ones." The essential point is that achievement is not expressed in abstract terms interpretable only normatively or by the teacher (if then). The parent can see what was required, what was submitted, and what evaluation the student gave the work.

Accountability in General

The tasks on which students work in cooperative small groups perhaps contain the skills, knowledge, synthesis, evaluation, and the like that define the learning framework. The term, "learning framework" is used because the boundaries are not fixed in advance. As with curriculum guidelines, minimum but not maximum targets are set. Unlike guidelines, the targets are expressed in terms of meaningful tasks that contain the skills and knowledge that now make up most guidelines. Accountability to the principal, board, and Ministry may be expressed in terms of targets set and targets attained, supplemented with reports of excellent work done over and above minimum expectations. The tasks should be listed, and the performance required, and then summaries of attainments provided that will be at least as informative as the current average percentages and test scores. Convincing all stakeholders of this is likely to take considerable time and effort. There are those who cannot live without tests and test scores. Unless schools everywhere begin to use the same tasks and report on them in the same way, there will still be demands to know where the schools stand in relation to one another. The most defensible educational solution will be for provinces or boards to survey achievement with a standard set of tasks, using light sampling to take a minimum of instructional time. If standard tests are the only means some in authority will accept, however, then they will have to be incorporated into the classroom work to give the students adequate practice. We will know we are making progress in education when there is widespread criticism of such practices, especially from those specializing in measurement, pointing out that they are educationally unsound and disruptive of teaching and learning.

References

- Airasian, P.W. (1988). Measurement driven instruction: A closer look. *Educational Measurement: Issues and Practice*, 7(4), 6-11.
- Archbald, D.A., & Newmann, F.M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Washington: National Association of Secondary School Principals.
- Bock, R.D., & Mislevy, R.J. (1988). Comprehensive educational assessment for the states: The duplex design. *Educational Evaluation and Policy Analyses*, 10(2), 89-105.

- Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook 1: Cognitive domain*. New York: McKay.
- Broadfoot, P., James, M., McMeeking, S., Nuttall, D., & Stierer, B. (1988). *Records of achievement: Report of the National Evaluation of Pilot Schemes*. London: Her Majesty's Stationery Office.
- Bussis, A. M., Chittenden, E.A., & Amarel, M. (1976). *Beyond surface curriculum—An interview study of teachers' understandings*. Boulder, CO: Westview Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Corbett, H.D., & Wilson, B.L. (1987). *Local effects of state testing programs: Preliminary data from a two-state study*. Unpublished report. Philadelphia: Research for Better Schools.
- E+M Newsletter, 45. (1988-89). Toronto: OISE Educational Evaluation Centre.
- Goldstein, H., & McLean, L. (1987, November). Assessment for the states—Possibilities and limitations: A critical look at the duplex design. *Evaluation Comment*, 16-17.
- Gorman, T. (1986) *The framework for the assessment of language*. Windsor, UK: NFER-NELSON.
- Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, 55(1), 23-46.
- Hargreaves, A. (in press). *Curriculum and assessment reform*. Milton Keynes, UK: Open University Press.
- McLean, L.D. (1988). Achievement measures made relevant to pedagogy. *McGill Journal of Education*, 23(3), 243-252.
- Metropolitan Toronto School Board. (1988). *Together we learn*. Toronto: Author.
- Radnor, H.A. (1988, September). *Coursework-how (if not when)—A study of a particular GCSE coursework scheme in practice*. Paper presented at the CEDAR conference, University of Warwick, UK.
- Swain, M. (1988, December). *Language measurement at the crossroads*. Paper presented at the 11th Annual Conference of the American Association of Applied Linguistics, New Orleans, LA.
- Thornton, G. (1986). *APU language testing 1979-1983: An independent appraisal of the findings*. London: Department of Education and Science.

ROSS E. TRAUB

Ontario Institute for Studies in Education

Assessment in the Classroom: What is the Role of Research?

The foregoing papers were reviewed from a perspective based on five propositions about classroom assessment. These refer to the nature of achievement, the information about students needed to interpret assessments of achievement, the tension that exists between the need to describe and the need to appraise, the distinction between formal and informal assessments, and the need for fairness.

This discussion of the papers commissioned for the Victoria Conference is divided into two main sections: The first consists of a perspective on classroom assessment in which several propositions about the nature of the enterprise are advanced. (I have chosen to substitute the phrase classroom assessment for classroom testing. Testing suggests particular kinds of activities, whereas assessment is less specific in connotation and is often used to encompass any reasonable means imaginable for collecting information about student achievement.) The second section contains comments on the papers and suggestions for future research.

Perspective

The perspective derives from five propositions:

Proposition I: Scholastic achievement is inferred from student behavior and from products of that behavior.

The evidence we obtain of a student's achievement derives from his or her performance of particular tasks on particular occasions. This evidence consists of an evaluation either of the performance of the task at the time it occurs (e.g., an adjudicator's rating of the performance of a composition on a musical instrument or a teacher's appraisal of the conduct of a laboratory experiment) or an evaluation of a product resulting from the task performance, this evaluation often being made some time after the performance (e.g., the teacher's appraisal of a student's answers to the questions in a written test). The evidence collected is, for most kinds of achievement, only a sample of the myriad kinds of evidence it would be possible to collect. So, although we refer to the performance or the products that result from the performance of particular tasks, our interest usually extends beyond particular performances of particular tasks to the performance of tasks similar to those that were attempted and perhaps also to the performance of tasks that are in some respects unlike those that were performed. From this

Ross Traub is a graduate of the University of Alberta (B.Ed., 1959; M.Ed., 1961) and of the Princeton University (Ph.D., 1964), and is on the faculty of the Ontario Institute for Studies in Education. His interests are in educational measurement and the impact of testing and assessment on education.

perspective, the high achieving student is one believed capable of behaving in the future with reference to familiar tasks, and possibly also with reference to unfamiliar tasks, in a way that will be judged acceptable or appropriate or noteworthy. An associated issue, then, is how widely the expectation of achievement for a student can be generalized: What limits should be imposed on the inference made from the sample of behavior and behavior products that served as the basis for the inference of achievement?

Proposition II: Achievement cannot be understood or interpreted without reference to the conation and affect of the actor, as well as her or his cognition.

The point of this proposition is that human behavior is a complex phenomenon, so making correct interpretations of the available evidence regarding a student's achievement is a hazardous exercise. According to the Oxford Shorter English Dictionary, cognition is "the action or faculty of knowing; knowledge, consciousness; a product of such an action." Conation is "endeavor; the faculty of volition and desire; also the product of this faculty." And affect is "mental disposition; desire, passion." Clearly, achievement is more than knowledge (cognition), for the behavior on which the assessment of achievement rests involves conative and affective components as well. In addition to knowing how to behave in ways that are appropriate and that serve to inform the observer, either through the behavior or a residue of that behavior, a student must also want to behave in a way that shows he or she possesses a particular kind of knowledge or ability. This may seem an obvious point, but we must be ever mindful of the fact that students who do not willingly perform tasks of the sort that provide evidence of knowledge and skill may be labeled disabled or underachievers for the wrong reason; namely, for the reason that they lack of knowledge or skill or ability, when instead they lack will or interest.

Proposition III: There is a tension in classroom assessment between describing and appraising student achievement.

Description may be characterized as the act of preparing a list or narrative summary of a student's behavior, including, perhaps, a collection of associated products, all of which, in the view of the person preparing the description, reflect achievement and support whatever inference is made of achievement. Appraisal goes at least one step beyond description and involves making judgments of good/bad, acceptable/unacceptable, strong/weak, high/low. This is necessarily done with reference to a set of norms or benchmarks, whether explicit or implicit. Assessments of achievement serve different purposes, but I cannot think of one that is wholly satisfied by complete reliance on description. The diagnostic purpose may come closest to being satisfied by description. The foremost consideration in diagnosis would seem to be description that reveals to the student and his or her mentor/coach/teacher that which is inadequate about a behavior or product, and that which needs to change in the student's cognition/conation/affect to alter the task performance in appropriate ways. Even here, however, there is implied contrast, between an actual behavior/product and a satisfactory or desired behavior/product. Thus good description is insufficient even for diagnosis. As Robert Wilson noted in the paper he prepared for this conference, description

is the main function of student assessment in the earliest grades of school. But even then, and certainly in later grades, it is expected that the teacher will judge whether a student's behavior and the products generated by that behavior satisfy expectations for the student's grade level. These expectations might be explicitly defined in a norms table, which is a summary of the way students representative of some well-defined population performed a standard set of tasks. More frequently, however, the student's behavior and the products of that behavior will be referred to expectations that the teacher has developed and internalized through experience, both with students and with subject matter or curriculum. Even the so-called criterion-referenced test involves a normative component whenever a student's performance on such a device is used to support a judgment of adequacy of achievement.

Proposition IV: In the act of appraising student achievement, a distinction can be drawn between informal and formal evaluations.

Informal evaluations of student achievement, as this notion is used here, are based on evidence obtained using tasks that are not standardized and for which norms are usually not available. Formal evaluations of student achievement are based on evidence obtained using standardized tasks and, typically, with reference to norms for judging the relative quality of the task performance. (Norms reflect variation in the performance of a standardized task by persons representing a clearly defined population.) A standardized task, analogous to a standardized test, is one designed for use by different persons in different places. It is intended that every use of a standardized task be governed by prescriptive rules for administration and scoring, these rules having been developed so that, when they are followed, the measurements or scores achieved by different persons in different places can be compared directly. Many, perhaps most, of the tasks used by teachers to assess student achievement are not standardized, hence they yield informal evaluations of achievement. Examples of formal evaluations include those based on student performance on province- or state-wide examinations in a subject, the norms being derived from the distribution of scores for all the students administered the examination.

Proposition V: Evaluations of student achievement that are used for selection or for determinations of pass/fail seem to demand, for the sake of fairness and equity, some use of standardized procedures and reference to norms.

The issue addressed in this proposition is two-sided. On the one hand is the matter of comparability of evidence of achievement from student to student within a class, and from class to class within a school, board, province, or country. A belief many in society hold is that an order of merit should be established among schoolchildren, an order that reflects their relative levels of achievement of more-or-less similar knowledge and skills. Now, there would be no need for an order of merit and for comparability of evidence (a) were there no need to make selection and pass-fail decisions affecting an individual's progress toward his or her goals in life, and (b) were there no expectation on the part of parents and society at large that assessments of how well a child is progressing in school would be provided in the normal course of events. Even recognizing these considerations, it seems

clear that comparability is not the end-all of achievement assessment. It is obviously desirable for students to develop in accordance with their different abilities and interests and personal goals. The question, when assessments of achievement are being made, is how to make fair comparisons among students who have attempted to achieve different things in school.

Teachers collect a variety of information on the achievements of students. Some of this information is informal and not comparable from one student to another. Consider, for example, the assessments made by means of projects, where each student undertakes a project of his or her own choice. Other information may be comparable from student to student within the same classroom, for example, scores on a test written by all students under more-or-less standard conditions. What was called formal evaluation in Proposition IV seems required for comparability of assessments across classes and schools. The results of formal evaluations can play a role in aligning the marking scales of informal evaluations, where different students are judged for their performances of different tasks. The difficult task for the teacher is to combine all the information, both formal and informal in the senses used here, into a composite that provides a basis for making fair comparisons of different students. The way in which all this information can or should be combined is a largely unresolved issue.

Some Comments on the Papers and Some Thoughts About Research

Six of the seven papers commissioned for this volume can be sorted into three loosely defined categories that relate directly to classroom assessment. The first category includes the reports of how student assessment is being done in British Columbia or Ontario or both. Here are the papers by David Bateson and Robert Wilson. These papers build on and otherwise tie into the results of other research conducted in Canada (e.g., McLean, 1985) and elsewhere. The second category includes the three papers in which procedures for changing, if not revolutionizing, classroom assessment are proposed. These were prepared by Dan Bachor, Clement Dassa, and Leslie McLean. Alone in category three is the contribution of Todd Rogers. His is a paper about student assessment and educational policy at the provincial level, with the issue being that of how provincial policy affects classroom practice in the domain of student assessment. The remaining paper by Philip Nagy and Rosalyn Moorhead lies outside a focus on classrooms and addresses an issue in principal leadership, albeit an issue rooted in classroom assessment. The research reported by Nagy and Moorhead is remarkable for the creative approaches they devised for summarizing qualitative, verbal reports.

Never Enough

The studies reported by Bateson and Wilson employed very different methodologies, very different samples, and were addressed to very different questions. Although both were essentially descriptive studies, Wilson's study differed from Bateson's in a way worthy of notice: Underlying Wilson's work was a theoretical framework, which informed the design of the study and the interpretation made of the results.

As interesting as descriptive research of the kinds reported by Bateson and Wilson is there never seems to be enough of it, or the right work seems not to have been done, to explain phenomena that arise in other contexts. For example, I read the Bateson and Wilson papers with the hope of finding clues to an explanation of the different rankings of British Columbia and Ontario 13-year-olds in the IAEP study, *A World of Differences* (Lapointe, Mead, & Phillips, 1989). One hypothesis is that British Columbia students have had greater experience with multiple-choice science tests like that used in the IAEP study than students in Ontario. I tried to confirm this in Wilson's report but was disappointed to find that science was not a subject of report in Wilson's Ontario sample at the 7th and 8th grade levels: the grades, in Ontario at least, where the vast majority of 13-year-olds are found. Of course, Wilson's study was not intended to yield widely generalizable results, so even if his study had included one or two 7th or 8th grade science classes, the evidence would have provided a far-from-conclusive test of the hypothesis. The foregoing example suggests a kind of research that may be needed: research into the effects of the mode of classroom evaluation on quality of student response to the tasks included in external assessments when those assessments employ, on the one hand, a mode frequently used in the classroom and, on the other hand, a mode infrequently used in the classroom. Many research questions of this kind could be posed, of course, so the matter of priorities must be addressed. I have no framework to offer for setting priorities. There is need for a thoroughgoing review of research on classroom assessment, with a view to constructing a framework on which to base priorities and funding decisions.

Advocates of Change

The papers by Bachor and McLean complement one another. Both call for changes in the way achievement is assessed in school classrooms. McLean sees the way ahead as that of collecting systematically the products that have resulted from student responses to "meaningful tasks." As I understand the term, a meaningful task is one that both student and teacher agree is relevant to a specified educational objective or kind of achievement. The evaluation of a student's response to a meaningful task would consist of a mark on a suitable scale of quality or merit. Bachor too, but for different reasons, sees merit in the idea of broadening the data base of classroom assessment. Bachor is addressing the situation in which students may not provide much evidence of achievement and are highly unlikely to be much interested in performing tasks on demand. In his scheme, the assessment data include observations of a sample of classroom performance, an analysis of a sample of work, a record of an interview of the student as he or she tackles a series of tasks, and a record of the student's response to a trial lesson.

The suggestions of Bachor and McLean are timely and appropriate. To assess a student's achievement is to make an inference based on evidence (the performance of tasks). There is no good reason to limit the evidential basis on which that inference rests; indeed, the narrower the evidential basis underlying an assessment of student achievement, the less generalizable the

assessment will be to the performance of other tasks, whether more or less similar. Performance of paper-and-pencil tests is altogether too narrow a base for assessing student achievement in the classroom. Inherent in the recommendations of Bachor and McLean, however, are problems that must be addressed. Most serious, perhaps, is that implementation of these recommendations might result in situations where the evidence in the record for one student is very different from the evidence in the record for another student. Such a lack of comparability would mean that direct comparisons of the achievement of one student against that of another, as in performance of the same task, would not be possible. Were this the case, such comparisons as were made would be done subjectively by the teacher, using whatever norms he or she had formed on the basis of experience for the different tasks and whatever subjective equating he or she had made of the different tasks and norms. Research is needed to provide a better description than we now possess of how teachers, operating independently, attempt to keep their marks in line with those given by other teachers of other students in the same grade or course, if indeed they attempt to do so at all. In his presentation, McLean passed lightly over the issue of assigning grades to cumulative records of achievement. (In a personal communication, McLean indicates that he has dealt with this issue in other reports.) How might it be ensured that a mark of A (or B or C) means more or less the same thing when assigned to different records of achievement? A solution, however imperfect, would be to ensure that every student's portfolio included several records of performance of exactly the same tasks. This information would provide a basis for adjusting the subjective and noncomparable scales that are used for assessing performances of tasks that are different from student to student. Time constraints on teachers being what they are, I imagine the written classroom test will continue to serve as the principal means of aligning the portfolio assessments of different students in the same class. In addition, of course, there is need either for the sharing of tests among teachers or for the occasional use of standardized tests so that teachers in different classrooms and schools have more or less comparable expectations of student achievement. Fairness and equity, if not society, demand no less.

A final point about the Bachor/McLean proposals: As desirable as a world without measurement error would be, such a world cannot be legislated into existence. The grades associated with cumulative records of achievement cannot be interpreted properly unless evidence of components of error variance is systematically collected. Thus the McLean and Bachor proposals include the germ of a research agenda. High on this agenda is research to identify the salient dimensions of variation in teacher judgments of portfolios or cumulative records of achievement. This might be accomplished using the multidimensional scaling approach employed by Glass (1967) in another context. Another item on the agenda is the conduct of generalizability studies (Brennan, 1983) to assess the magnitude of the components of variance associated with the marks teachers assign student records. (An alternative approach, suggested by McLean in a personal communication, is to create, if possible, a theory of measurement that applies in situations where the data are not scores on some continuous scale but, instead, are assignments to one

of a few categories, these assignments being made under the conditions that the number of students is small, as in the typical classroom, and the assessments, that is the assignments to categories, are made frequently.)

Dassa's approach to reforming classroom assessment is very different from that recommended by Bachor and McLean. Disillusioned from the experience of prescribing a method of reform based on a particular model, in this case the mastery learning model of formative evaluation, Dassa has taken the approach of meeting with teachers and, as I understand it, working with them to identify and solve their assessment problems. The cornerstone of Dassa's approach is to gain the trust of each teacher, thus to be in a position to bring his or her expertise to bear on their problems. Dassa's goal, it seems, is to bring the assessment approaches of different teachers as closely into line with a formative evaluation model as possible. This approach is reminiscent of that taken by Sarason (1971) and his colleagues two decades ago as they endeavored to change practice in a school near New Haven, Connecticut. In the course of his work, Dassa expects to define a "research paradigm centered on the classroom." I can only wish him well in this pursuit and eagerly look forward to reading an evaluation of its success.

Policy Considerations

Rogers' paper gives us an excellent account of power politics as played in the assessment arena. Clearly, governments that see a need for increased external surveillance of student achievement, and that have public support for such surveillance, will introduce policies and associated practices that are likely to affect classroom assessment of student achievement. What is needed in this connection are studies of assessment policies and practices at all levels from the province (state) to the school board to the school to the classroom (teacher), both before and after public policy and practice have been changed. This will make it possible to associate changes at the different levels with changes at the provincial (state) level. The resulting information will enlighten any debate as to the possible effects of provincial policy and practice on the day-to-day experiences of students. It is gratifying to know that British Columbia is funding studies of the possible effects on students, teachers, and teaching practices of the Grade 12 examination program and the British Columbia assessment program. It is to be hoped that government officials in other jurisdictions will see the need for, and fund, this kind of research. Although not a research agenda per se, members of the measurement community should press for such studies.

References

- Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City: American College Testing Program.
- Glass, G.V. (1967). Factors in teachers' perceptions of students. *Journal of Educational Measurement*, 4(2), 87-93.
- Lapointe, A.E., Mead, N.A., & Phillips, G.W. (1989). *A world of differences*. Princeton, NJ: Educational Testing Service.
- McLean, L.D. (1985). *The craft of student evaluation in Canada*. Toronto: Canadian Education Association.
- Sarason, S.B. (1971). *The culture of the school and the problem of change*. Boston: Allyn & Bacon.

RICHARD J. STIGGINS

Northwest Regional Educational Laboratory, Portland, Oregon

Toward a Relevant Classroom Assessment Research Agenda

This commentary underlies the need to pursue measurement research in the classroom setting. We need to describe assessment processes as they occur and understand their effects. From the papers which support the new research agenda, implications for practice are drawn.

Considered collectively, the papers presented at this conference provide a broad perspective on the state of classroom assessment research, development, and training. As I read the papers, heard the presentations, and participated in discussions, I tried to identify emerging themes in this growing field of study. In my summary, I would like to share some of my thoughts regarding those themes, focusing on the goals or objectives that appear to be guiding our research, the research methods we are using, some of the results being attained, and emerging implications for future work.

Research Goals and Objectives

Clearly as a community of scholars and practitioners, we have made a complete transition from caring totally about standardized tests to caring far more about better classroom assessments and better classroom-level decision making. These papers attest to this fact.

Further, it is clear and appropriate that our most important research objective has been to describe the status of classroom assessment methods. We are describing many aspects of that status, including the various assessment methods teachers use across different student and subject matter contexts. We are documenting elementary and secondary assessment differences, as well as the unique dimensions of assessment across such content areas as mathematics, writing, speaking, and science. Papers by Bateson, Wilson, and Bachor represent examples of such descriptions. Through this work, we are showing that assessment is not assessment. Rather, as the nature of the achievement target varies across grades and subjects, so do the appropriate assessment methods. As a result, we now know that teachers face complex task demands as they strive to track the achievement of their students.

Further, in our research at the Northwest Regional Educational Laboratory (NWREL), we have joined forces with those striving to document the

Richard Stiggins is the director of the Center for Classroom Assessment. His interests center on applied classroom assessment.

nature and quality of the training and technical assistance provided to teachers to help them meet those task demands. Documentary evidence is now in hand to corroborate what we have always sensed but rarely verbalized: that training is totally inadequate if it is presented at all, and technical backup in assessment is virtually nonexistent for teachers.

As we set our objectives for the future, we should continue to describe the nature and quality of classroom assessment processes in as much rich detail as we can. But in addition, let us follow the path spelled out by Dassa in his paper and strive to understand the effects or implications of the various assessment processes teachers use on student development. In my opinion, this represents one of the next two critical classroom assessment research frontiers: We must strive to understand how our assessments impact students, who are, after all, the most important decision makers relying on the results of those assessments. Students set expectations and decide how to relate to school on the basis of data provided to them by their teachers. If they receive sound data, they make sound decisions. If they receive unsound data, they make unsound decisions. They and they alone are the beneficiaries or victims of classroom assessment. Yet we know so little about how they regard, evaluate, and act on the data they receive.

The second great research frontier we face is that of teacher and administrator training in assessment. How can such a critical component in the professional preparation of teachers have been so completely ignored or mismanaged? Researchers can investigate and understand the reasons for the lack of relevant training. We can find ways to develop and deliver relevant training. We understand the meaning of quality assessment. We can demystify our technical concepts. The standards of acceptable training are not mysterious. Our research task is to find meaningful ways to communicate those standards to educators who must strive to meet them.

In sum, these papers confirm that we are aspiring to appropriate goals, as we continue to describe the complexities of classroom assessment processes and environments. In addition, we must work to understand the critical effects of assessment on students and we must begin to sort out effective strategies for rendering assessment training relevant and helpful for practitioners.

Research Methods

The research methods presented in these papers are important because they represent promising departures from the traditional array of methods used in psychometric research. The idea, for example, of presenting educators with a specific assessment problem and observing how they attack it, as Nagy and Moorhead have done, opens a range of possibilities for describing and understanding the current state of classroom assessment. The examination of special education assessment methods in terms of their implications for the general student population, as Bachor has done, demonstrates a kind of professional reflection that can enhance the quality of classroom assessment. And finally, the experimental method demonstrated by Dassa reminds us of the continuing value of effects research.

All of these research methods hold the promise of expanding our wisdom in the arena of classroom assessment. At NWREL, we have relied on all of these methods to build our knowledge base. In fact, we feel we experienced a major breakthrough in our understanding of the task demands of classroom assessment when we became ethnographers and entered classrooms as participant observers with the goal of understanding the assessment experience from teacher and student perspectives. As we pursue our goal of continuing to describe and understand day-to-day assessment processes, we continue to rely on this methodology. Our next research goal is to find cases of success—outstanding classroom assessment environments—so we can contrast them to less successful classroom assessment environments.

As we look to the future, I believe we attain maximum success in understanding the nature and quality of classroom assessment if we rely on research methods that emphasize depth of information more than generalizability of results. That means small samples viewed with powerful microscopes. However, as we strive to find ways to enhance the quality of assessment, we might return to more classical research paradigms. We can evaluate various training methods via experimental and quasi-experimental approaches to find those that have maximum impact. As we conduct these studies, we must examine the impact of training on several dependent variables: trainee knowledge of assessment practice, implementation of sound practice, confidence in assessment results, and impact on student learning.

Research Results

Our various research results are coalescing around consistent and important themes: (a) the day-to-day assessment of student achievement in many classrooms is complex and demanding and (b) is not producing the high-quality data for decision making that we would like, because (c) those teachers and administrators who serve in the front lines of education have not been provided basic training in classroom assessment methodology. We know what assessment methods teachers use and the extent of their use. We know that teachers rely on far more than objective paper-and-pencil tests. We understand many of the assessment quality control problems that arise in classrooms and we know about the frequency of their occurrence. As McLean has told us in his paper, we know a great deal about the “craft knowledge” that exists in classrooms about the assessment process. Further, we know that that craft knowledge does not reflect the sense of assessment captured in the traditional standardized test.

In short, we understand a great deal about classroom assessment, even if we have a great deal yet to learn. These papers suggest that we should be encouraged by a decade of research that has taken us this far. The next decade will carry us far beyond. We know the quality of classroom assessment can improve.

Implications for the Future

As we work for that improvement, here are some key perspectives we need to keep in mind:

1. *Policy makers will play a critical role in improving classroom assessment.* The assessment community knows that teachers and administrators are in desperate need of assessment training; teachers and administrators know they need assessment training, but policy makers often are not aware of this gap in professional preparation. As a result, they make policy, as Rogers reveals in his paper, on the assumption that practitioners can handle all relevant assessment issues. For example, they approve plans for a province-wide ungraded primary on the assumption that teachers are prepared to handle the considerable assessment challenge that accompanies the need to individualize instruction. If this assumption is invalid (and we know it is for many teachers), such a policy may result in a program doomed to fail before it begins.

Policy makers need to (a) become literate in the basic principles of sound assessment themselves, and (b) become aware of inadequate assessment training provided to the educators who will implement their policies. The assessment community must take the lead in training policy makers and sensitizing them to these issues.

2. *Classroom assessment is a domain of professional practice that involves a wide variety of assessment methods.* Many in our society believe that most, if not all, of our valued educational outcomes can be translated into multiple-choice test items. We must help them overcome this misunderstanding.

As educators, we help students to hit many different kinds of achievement targets: content, thinking skills, behaviors, and products. Some targets can be translated into objective test items; others cannot. Those who educate and those who view education as interested outsiders need to understand that the assessment of the full range of valued outcomes requires the full range of assessment tools: paper-and-pencil tests, performance assessments (observation of and judgment about behaviors and products), and personal communication with students. Each represents a valuable assessment tool when used well in the proper decision context. All must be made aware of this fact.

3. *In the past, practicing educators have either not been motivated or not been given the opportunity to learn about assessment methodology.* This is no longer acceptable. We need to conduct the research needed to understand how to motivate (a) teacher training institutions to provide relevant assessment training, and (b) teachers and administrators to take advantage of training opportunities. We have at least two strategies available for reaching these goals: force or finesse. We can force participation in training through power of law. Teacher training institutions can be made to adhere to revised standards of acceptable assessment training. Teachers can be made to participate in training through revised program requirements. They can be made to pass rigorous certification tests revised to cover assessment methodology.

Or we can make available to educators relevant and helpful training and encourage participation to grow through spreading interest in new approaches to classroom assessment. In this case, all training would be offered and

taken on a voluntary basis. However, it would be made to be very attractive training.

We have in hand both the background information and teaching strategies to provide relevant, helpful assessment training. We must now research the best way to motivate educators (and noneducators alike) to take advantage of that training. If finesse will work, it is the best option. But if finesse is insufficient, the force of law may be needed.

4. *Traditional measurement courses, per se, won't meet the training needs of all teachers or all contexts.* We need alternative modes for presenting training. We need training strategies and modes of presentation that fit both the preservice training context and the huge inservice training need. Written training materials, video training, audiocassettes, and other available instructional technology can be used profitably to communicate keys to effective assessment using both extended presentations and more efficient brief contacts. We must explore all such options.

5. *Existing technology promises to revolutionize the gathering and use of information in the classroom, if we learn to apply it.* Research is needed on strategies for using optical scanners, database management systems, word processing, gradebook software, video presentations of information, computer item banks, and other advanced technology to ease the information management load on teachers and administrators. The technology exists; we must learn to use it to advantage. If we wish to "sell" sound classroom assessment to teachers as a process that can make their jobs faster, easier, and better in that order, the use of available technology is an excellent way to begin.

However, we must take great care not to offer technology as a substitute for understanding the essentials of sound assessment. For instance, we cannot profitably offer banks of previously-developed test items to teachers as a substitute for teaching the users how to write and select good test items.

Without the basic assessment training, they will not know how to use the item bank to maximum advantage. The bank will come to represent just another variation on the same old test construction hassle and they will not use it. Only those who understand sound assessment can use advanced technology to improve assessment quality and efficiency.

6. *Excellent research is being conducted on classroom assessment issues, yet it is not being published and disseminated for critical review, evaluation, and synthesis.* We must develop more and better ways to include reports of classroom assessment research in scholarly journals. The Surrey, British Columbia School District's journal *Research Forum* (a journal devoted to educational practice and theory) represents an excellent example of one attempt to find new ways to communicate. Other districts and professional associations might follow this lead. In addition, research on classroom assessment issues, with its many and varied research methods, needs to be elevated to an appropriate level of acceptability for university faculty tenure and promotion evaluations and for dissertation and thesis research in graduate programs. Through these steps and others like them we can both encourage

and spread word of important classroom assessment research and development.

Conclusion

The lessons of history tell us that when assessment research, development, and training agendas mix topics related to large-scale, high-stakes tests with topics related to classroom assessment, the much less visible and less politically important classroom assessment issues almost always lose. High-stakes tests are like sponges, drinking up the resources.

The time has come to separate the two. They represent different disciplines, resting on different underlying assumptions and paradigms. They serve different decisions and different decision makers. The methods of classroom assessment are more diverse, complex, and demanding than those of high-stakes testing. The time has come to separate them. The effect of doing so has been demonstrated by this challenging, stimulating, and exciting conference. The focus was narrowed. We discussed not the improvement of all of school testing, but the improvement of classroom assessment. The result is a conference that suggests specific action. Let us begin.

Curriculum and Teaching is a new international review journal which will publish original research and contributions dealing with major and current issues and trends in curriculum theory and practice. The object of the journal is to advance the study of curriculum and teaching, the development of which might improve teaching and pedagogy. *CT* will provide a new and impartial forum for teachers and scholars throughout the world, working in the area of curriculum studies and instruction. The journal will examine and evaluate current issues in curriculum, instruction and educational theory — from a broad, balanced and comparative perspective. The journal, by focusing on both curriculum theory and teaching, hopes to appeal to many teachers, administrators, researchers and teacher educators throughout the world. The journal will devote its attention to curriculum design and development, evaluation, curriculum models, comparative studies in curriculum, innovation, and policy, planning and educational administration. Each issue of the journal will contain major articles, research reports, comparative case studies, and book review and notes.

Contributions (normally no more than 6,000 words, typed double spaced in 3 copies) should be addressed to the Editor, *Curriculum and Teaching*, James Nicholas Publishers, P.O. Box 244, Albert Park, Australia, 3206. In order to ensure that articles are of highest quality all contributions are submitted to at least two independent referees before acceptance for publication.

The journal will appear twice a year, in April and October.
These two issues will constitute one volume.
Volume 3, No. 1-2, 1988.
ISSN 0726-416X

Orders (including remittances relating to back issues) should be sent to the publishers: James Nicholas Publishers, P.O. Box 244, Albert Park, Australia, 3206.

EDITORIAL BOARD

Professor Michael Apple *University of Wisconsin-Madison*
Professor Harold Bergsma *University of New Mexico*
Professor John Eggleston *University of Keele*
Professor Henry Giroux *Miami University*
Professor Allan Glatthorn *University of Pennsylvania*
Professor Kenneth Howey *University of Minnesota*
Professor Anne Jefferson *University of Manitoba*
Professor Robert McNergney *University of Virginia*
Professor Thomas Popkewitz *University of Wisconsin-Madison*
Professor George Posner *Cornell University*
Dr. David Reynolds *University College, Cardiff*
Professor J. A. Sweeney *University of Texas at Austin*
Professor Edgar Stones Editor, *Journal of Education for Teaching*

Professor P. H. Taylor *University of Birmingham*

CONSULTING EDITORS

Professor Charles Burgess *University of Washington*
Professor Elliot Eisner *Stanford University*
Professor Paul Hirst *University of Cambridge*
Professor Daniel Kallos *University of Lund*
Professor Arie Lewy *University of Tel Aviv*
Professor Kenneth Sirotnik *University of California, Los Angeles*
Professor Malcolm Skilbeck *University of London, Institute of Education*
Professor Margaret Sutherland *University of Leeds*
Professor Laurel Tanner *Temple University*
Dr. Joseph Zajda (Editor) *Institute of Catholic Education*
MANAGING EDITOR
Rea Zajda *James Nicholas Publishers*

RECENT ARTICLES INCLUDE

Laurel N. Tanner *Education and the Idea of Progress: 1884-1984*
Thomas Popkewitz *Curriculum Reforms, Educational Rationality and Change*
Henry Giroux *Teaching and the Role of the Resisting Intellectual*
Michael Apple *The Political Economy of Text Publishing*
Anne Jefferson *The Management of Conflict Associated with the Implementation of Change*
Ken Howey, Nancy Zimpher *New Curriculum Directions in the Education of Teachers*
Alan Tom *Curriculum as a Moral Craft*
Robert McNergney Martin *Aylesworth/Evaluating Pre-Service Teachers*
Joseph Zajda *Moral Curriculum in Soviet Schools Today*

CURRICULUM AND TEACHING

ORDER FORM

Please enter my subscription to *Curriculum and Teaching*

I enclose US\$

Name

Address

.....

.....

Signed Date.....

For Contributors

Please prepare manuscripts in accordance with these guidelines.

Format

- Manuscript in triplicate.
- Typewritten or printed, double-spaced throughout.
- Abstract of approximately 100 words on a separate sheet.
- Manuscript not to exceed 6,500 words excluding graphics.
- Author's name and affiliation on a separate title page; title only on first page of manuscript to ensure anonymity in the review process.

Referencing

- Sources in parentheses after each reference (direct or otherwise), giving author's name (unless mentioned in text), year of publication, and page number(s).
- Direct quotes of 40 words or less in double quotation marks incorporated into text; those longer than 40 words in indented block format.
- All sources listed alphabetically at the end of the manuscript under the heading *References* using the APA style.
- Explanatory notes, numbered consecutively and marked in the text with superscript numerals, may appear before the References under the heading *Notes*; any citations in notes follow the same format as other references.

Graphics

- Each table and figure on a separate page at the end of the manuscript.
- Figures, charts, and diagrams in camera-ready form.
- Number tables in Arabic numerals and center over each heading, e.g., *Table 1*.
- Number figures in Arabic numerals before the caption at foot of each graph, e.g., *Figure 1: ...*
- Indicate placement of figures and tables in text, i.e., *Insert Table 3 here*.

Style

- *AJER's* editorial style conforms closely to the *Publication Manual of the American Psychological Association*.
- Editorial changes may be made to manuscripts.
- For spelling, consult *Webster's New Collegiate Dictionary*; spelling in quoted material remains as in the original.

Diskettes

If manuscript is accepted for publication and has been prepared using a Macintosh, IBM, or IBM compatible microcomputer, please send diskette. Microsoft Word, Multi-Mate, WordPerfect, WordStar, Xerox Writer, XyWrite, and DisplayWrite 3 **only** are acceptable.

Important: Please indicate program used.

Submission to *AJER* implies the manuscript is not being considered for publication elsewhere.

AJER is indexed in the *Alberta Education Index*, *Canadian Education Index*, *Current Contents/Social and Behavioral Sciences*, *ERIC/Current Index to Journals in Education*, and *Social Science Citation Index*; appropriate articles are abstracted in *Cognitive Development Abstracts*, *Educational Administration Abstracts*, *Educational Technology Abstracts*, *Family Abstracts*, *Language Behavior Abstracts*, *Multicultural Education Abstracts*, *Psychological Abstracts*, *Research into Higher Education Abstracts*, *School Organization and Management Abstracts*, and *Sociology of Education Abstracts*, and reviewed by the *Journal for Research in Mathematics Education*.

